Computers & Security 145 (2024) 103988

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Adversarial Machine Learning in Industry: A Systematic Literature Review

Felix Viktor Jedrzejewski [a,*], Lukas Thode [a], Jannik Fischbach [b,c], Tony Gorschek [a,c], Daniel Mendez [a,c], Niklas Lavesson [a]

[a] Blekinge Institute of Technology, Valhallavägen 1, 371 41 Karlskrona, Sweden
[b] Netlight Consulting GmbH, Sternstraße 5, 80538 Munich, Germany
[c] fortiss GmbH, Guerickestraße 25, 80805 Munich, Germany

ARTICLE INFO

Dataset link: https://zenodo.org/records/10654103

Keywords:
Adversarial machine learning
Industry
Rigor
Relevance
State of evidence

ABSTRACT

Adversarial Machine Learning (AML) discusses the act of attacking and defending Machine Learning (ML) Models, an essential building block of Artificial Intelligence (AI). ML is applied in many software-intensive products and services and introduces new opportunities and security challenges. AI and ML will gain even more attention from the industry in the future, but threats caused by already-discovered attacks specifically targeting ML models are either overseen, ignored, or mishandled. Current AML research investigates attack and defense scenarios for ML in different industrial settings with a varying degree of maturity with regard to academic rigor and practical relevance. However, to the best of our knowledge, a synthesis of the state of academic rigor and practical relevance is missing. This literature study reviews studies in the area of AML in the context of industry, measuring and analyzing each study's rigor and relevance scores. Overall, all studies scored a high rigor score and a low relevance score, indicating that the studies are thoroughly designed and documented but miss the opportunity to include touch points relatable for practitioners.

1. Introduction

Machine Learning (ML), a subdomain of Artificial Intelligence (AI), focuses on adapting to new circumstances and detecting and applying patterns from experience in the form of data to solve a given problem (Russell and Norvig, 2010). Nowadays, ML is used in many software-intensive products and services in different security-critical industry domains, ranging from manufacturing (Bertolini et al., 2021) to medicine (Terranova et al., 2021). This trend of ever-increasing involvement of ML-based services in almost every part of our society also increases risks that emerge from the underlying ML models.

Biggio and Roli (2018) concluded that ML was originally designed to solve closed-world problems without defense against external and potentially malicious interactions. A tremendous switch of ML approaches to open-world problems increased the attack vector of ML models applied in products and services. Adversarial Machine Learning (AML) studies malicious third-party actions against ML models (Sadeghi et al., 2020) to force them to either misclassify input (Goodfellow et al., 2015), decrease their prediction accuracy (Steinhardt et al., 2017), extract their training data (Fredrikson et al., 2015), or steal them (Reith et al., 2019). In summary, AML focuses on exploiting ML-specific vulnerabilities and developing defense mechanisms to mitigate the exploitation of ML vulnerabilities (Biggio and Roli, 2018; Cinà et al., 2023). This leads to tremendous consequences observable in real-world incidents for software relying on ML model predictions, such as Anti-Malware scanners misclassifying ransomware[1] or Shanghai government tax office's facial recognition systems granting 77 million dollars of taxpayer money to criminals.[2] It is imperative to be aware of the potential attacks when incorporating an ML model into a product to avoid additional risks emerging from it.

After almost a decade of research in adversarial machine learning, counting over 2300 studies, organizations lack effective approaches to applying state-of-the-art defense techniques to defend their ML models, indicating a gap between academia and industry (Anderson, 2021). This is even further stressed through studies showing that unsystematic and low-effort attack approaches on machine learning systems yield concerning results (Apruzzese et al., 2022). Additionally, the researchers describe the setup of a crawled AI incident database by Tidjon and Khomh (2022) and analyzing its content, that many incidents lacked any form of tactic or heuristics. Moreover, one of the most common attack entry points is Reconnaissance. Attackers investigate publicly available materials, such as Adversarial Machine

* Corresponding author.
E-mail address: felix.jedrzejewski@bth.se (F.V. Jedrzejewski).

[1] https://atlas.mitre.org/studies/AML.CS0002/.
[2] https://atlas.mitre.org/studies/AML.CS0004/.

https://doi.org/10.1016/j.cose.2024.103988
Received 14 February 2024; Received in revised form 28 June 2024; Accepted 2 July 2024
Available online 17 July 2024

Learning (AML) studies, repositories, and pre-trained models, to facilitate attacks. AI incidents are not rare anymore, and the website AIIncidentDatabase provides the opportunity to report and read up about such cases. However, this database covers just a few successful adversarial ML attacks. Still, the work of Apruzzese et al. (2022) shows that cybersecurity companies face cases where ML models are the main target. This means that attackers have already launched adversarial ML attacks, and so far, either no major incident has been published because affected companies do not have a defense strategy to avoid further attacks or they unaware about ongoing attacks on their systems. Therefore, all ML-based services must protect their ML models to avoid the negative consequences caused by the exploitation of their vulnerabilities.

Multiple surveys interviewing ML practitioners have been published in the last two years, with titles already underlining the unawareness and current attitude towards adversarial machine learning, such as "I Never Thought About Securing My Machine Learning Systems"(Boenisch et al., 2021) or "Why do so?"(Grosse et al., 2022). A study by Apruzzese et al. (2022) reports that academics develop solutions with little to no practical value, going through top-ranked security conferences.

In this study, we report on a systematic literature review that investigates the state of reported evidence and available approaches through the lens of rigor and relevance. Rigor and relevance are metrics introduced by Ivarsson and Gorschek (2011) to measure and evaluate the transfer and dissemination of research results of a certain research domain in industry. Rigor measures and evaluates the extent and detail a study reported and obtained its results. Relevance reflects the degree of realism a study design adopted to mimic an industrial setting (Ivarsson and Gorschek, 2011). Our study will analyze the industrial applicability of the peer-reviewed literature through the lens of practitioners, which has been overlooked so far. The suggestions proposed in this study will motivate and contribute to improving problem-driven research and industrial adoption, forming more collaborations between academia and industry. A collaboration between academia and industry will be mutually beneficial and help identify and solve security-related challenges when using ML systems.

In particular, we make the following contributions:

- We report on a mapping study on contemporary literature on prominent attacks addressed by Adversarial Machine Learning research
- We compare the Adversarial Machine Learning research landscape with needs expressed by industrial practitioners
- We analyze proposed solutions based on Rigor and relevance

The remainder of the article is organized as follows: in Section 2, we provide background information about AML in the context of industry: in Section 3, we summarize the related work and how this study relates to it; in Section 4, we illustrate the methodology we used to carry out our literature review; in Section 5, we report the results of the literature review; in Section 6, we discuss the results of this study with related literature; in Section 7, we discuss the results and the threats to validity; in Section 8, we conclude the article by summarizing the key findings and suggesting future research directions.

2. Background

This section will explain the term Adversarial Machine Learning and the underlying attacks relevant to this study. Furthermore, we will explain the relationship between AML and zero-day attacks. Finally, we summarize the AML concerns raised by industry.

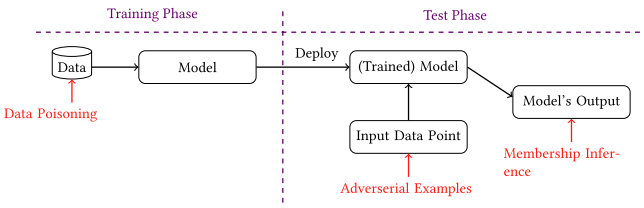

Fig. 1. AML attacks and their occurrence illustration.

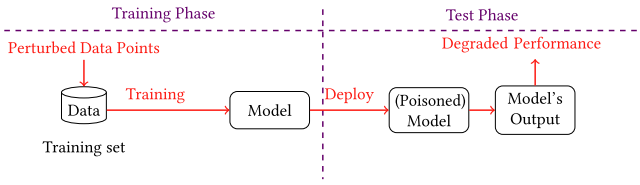

Fig. 2. Data poisoning concept illustration.

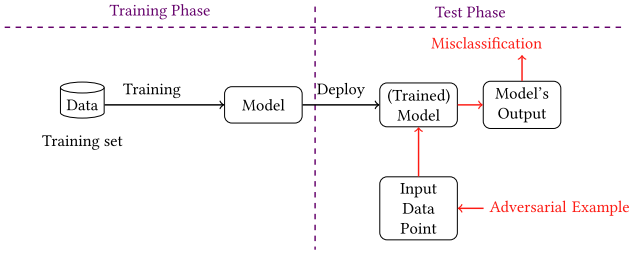

Fig. 3. Adversarial examples concept illustration.

2.1. Adversarial machine learning

Adversarial Machine Learning entails research investigating attacks targeting ML models and how to defend against them (Biggio and Roli, 2018; Barreno et al., 2006). We map the extracted challenges related to ML model security to the attack overview targeting machine learning models suggested by Bieringer et al. (2022). Attacks on ML models either occur during the training or testing phase. Fig. 1 illustrates the attacks relevant to this study.

Data poisoning. In the training phase, a malicious third party can conduct a Data Poisoning attack if it can access the training data set. Data Poisoning denotes the intentional injection of maliciously perturbed data points into the training set of a model to reduce its accuracy (Steinhardt et al., 2017) as displayed in Fig. 2.

After the training phase, the ML model is deployed in an environment where it receives input data and generates predictions. This is commonly referred to as the test phase. A malicious third party can attack the deployed model with Adversarial Examples (Goodfellow et al., 2015) or conduct a Membership Inference Attack (Shokri et al., 2017).

Adversarial examples. Adversarial Examples represent specifically crafted data points forged to force a deployed ML model to misbehave or misclassify given data (Goodfellow et al., 2015) as displayed in Fig. 3.

Membership inference. Membership Inference describes the process where an attacker succeeds in extracting training data points from a deployed model or determining if a data point has been used during the training phase of a given deployed model based on the model's output, violating model and training data privacy (Shokri et al., 2017).

Relationship between zero-day attacks and AML. A zero day attack entails the exploit of a vulnerability which was unknown to the public, hence, developers had "zero days" to patch the undisclosed vulnerability to mitigate the exploit (Bilge and Dumitraş, 2012). AML investigates vulnerabilities specific to ML models that are transferable to other ML models (Papernot et al., 2016; Suciu et al., 2018). Based on the novelty and transferability of AML attacks on other ML models, AML can provide implicit zero-day attacks because some AML attacks are impossible to mitigate systematically yet.

2.2. AML concerns raised by industry

Based on interviews with ML stakeholders from different companies, researchers concluded that Data Poisoning attacks occurred in an industrial setting (Grosse et al., 2022), highlighting the risk of incorporated ML systems. Data Poisoning and Model Stealth attacks are perceived as more relevant for practitioners in comparison to other kinds of AML attacks (Kumar et al., 2020). Furthermore, ML stakeholders point out challenges towards AML referring to awareness, risk assessment, and defense implementations (Mink et al., 2023).

3. Related work

We start describing the development of Adversarial Machine Learning by pointing out numerous secondary studies elaborating on the topics consisting of or referring to the terms "Machine Learning" and "Security" in different application areas. Next, we move into studies addressing AML and the developments within the domain and its intersection with industry, dissecting the causality of the gap between academia and industry and listing current mitigation approaches.

3.1. AML literature studies

A large number of literature studies are investigating machine learning and its security aspect in general. Interestingly, many studies cover Adversarial Machine Learning, indicating a general interest in this research field.

Wu (2020) investigates in a survey the role of noise in the training and testing data collected in Internet of Things (IoT) devices and then used for adversarial examples. The authors highlight emerging challenges and issues, among which the absence of a reliability layer in the architecture of neural networks stands out as a significant concern. Another challenge is a lack of research efforts for solution approaches for dealing with zero-day vulnerabilities. A robust testing framework to ease the test case generation process and evaluate an ML model's maturity concludes the highlights of the challenges. The article concerns theoretical aspects and fails to mention any collaboration with industry, which is crucial for practical applications. Although the article suggests more research for zero-day vulnerabilities and robust testing, it fails to emphasize that industry involvement would be beneficial and essential to tackling these challenges.

In another survey by Liu et al. (2018), four categories of defense techniques for machine learning are presented from two aspects (training and inferring). These are security assessment mechanisms, countermeasures in the training and testing phase, data security and privacy. As machine learning is used in numerous practical applications on a wide range, it simultaneously increases the attack vectors related to ML-specific security concerns. Liu mentions that existing research has produced a diverse range of literature studies that concentrate on the analysis of defense mechanisms for machine learning from various perspectives and under different circumstances. Moreover, it is pointed out that introducing security measures to machine learning algorithms would lead to an overhead, resulting in slowing down the performance, which the industry will not be ready to make. The survey concludes that new threats to machine learning are continuously emerging, which stresses the call to action, especially in the industry.

A survey by Li (2018) shows threats associated with artificial intelligence (AI) and an analysis of various defense strategies. Correspondingly, aspects of developing a safe AI in a distributed machine learning or deep learning environment are covered, and discuss a variety of attacks on AI in an adversarial environment. The survey highlights a research gap in the involvement of industry in future research endeavors, which is crucial for discovering efficient defense strategies. Last, Li (2018) emphasizes that academia and research institutions need to focus on the development of security protection solutions for AI.

Machine learning methods are highly vulnerable during the training and inference phase when exposed to attacks in an adversarial environment, stats a survey by Wang et al. (2019). The survey delivers a compact overview of the security traits of machine learning algorithms when operating under adversarial settings. In this more theoretically focused survey, which lacks the focus on industry fundamental solutions for defense systems against adversarial, examples of deep learning methods are provided. Moreover, Wang et al. (2019) mentions that the evaluation of ML methods under adversarial settings is an area that should receive more attention from the research community as the fundamental solution to input perturbations resulting in considerable changes in the output has not yet been found.

Qiu et al. (2019) present in a review a summary of the recent research for adversarial attacks and defense technologies in the field of deep learning. Here, defense strategies are divided into three categories: modifying data, modifying models, and using auxiliary tools. Further, this review mentions two research gaps. The first gap is that black-box attacks, for example, adversarial examples, model inversion, and model theft, have better applicability in real-world scenarios. Concluding Qiu et al. state that improving the robustness of AI systems against adversarial attacks is imperative within the AI development process, and attacks in the training stage are seldom in the real world. Additionally, the authors emphasize further research on the attack technology to identify defense strategies that can mitigate multiple types of attacks. Currently, the available defense mechanisms are inflexible and focus only on specific types of attacks. A more holistic defense approach capable of multiple types of attacks is needed.

Biggio and Roli (2018) shows a comprehensive overview of Adversarial Machine Learning over the last ten years, including guidelines on how to assess and improve the security of machine learning. As a research gap, guaranteeing trustworthy machine learning predictions based on uncontrollable and unforeseen user input is a task in the intersection of adversarial machine learning, robust AI, the interpretability of machine learning, and industry. The goal is to maximize research possibilities and enhance the efficiency of the solutions, as well as to assess and guarantee the security level of the ML solutions applied in products we use on a daily basis. Originally, machine learning was designed to handle closed-world problems. However, this has changed over the last years, causing an alternation of requirements for machine learning since machine learning models now face unforeseen inputs from the open world provided by uncontrollable users. Often, machine learning practitioners are unaware of AML, which implies that most systems are vulnerable to known unknowns. This can be utilized by attackers to exploit machine learning models and the corresponding systems that are embedded in them.

Dedicated subtopics of adversarial machine learning are also examined in various literature studies. For example, Yuan et al. (2019) shows a set of attack and defense techniques specifically designed for deep neural networks. A more detailed overview of adversarial examples in deep neural networks can be found in the work of Zhang and Li (2020). The result is an overview of state-of-the-art methods for adversarial example generation. Another study by Sadeghi et al. (2020) introduces and applies a new taxonomy and a research impact metric on 250 AML studies. The results are used to evaluate the research impact of every AML study and illustrate the security arms race, a term describing the continuous breaking of defense mechanisms protecting ML models caused by novel attacks targeting them. Sadeghi et al. (2020)'s study

touches on the context of industrial relevance and points out the lack of practical studies. Data Poisoning and Backdoor are surveyed by Cinà et al. (2023) and Li et al. (2022). They provide an overview of the research landscape for each attack, analyzing current trends and pointing out research gaps.

Overall, Apruzzese et al. (2022) analyzes 88 studies from prominent security venues and concludes that most studies fail to complement the industry with their results because they solve issues in experimental environments irrelevant to the industry. To foster the collaboration between industry and academia, Apruzzese et al. (2022) propose to (1) develop threat models for ML systems, not just ML models, (2) focus more on the economics of cybersecurity, (3) solve AML issues in a real-world ML system instead of a ML model in a laboratory environment.

Most surveys about AML are theoretical because the studies in the field target ML models handling close-world problems. Therefore, the possibility of an attack launched by an external malicious third party was, given at that time, highly unlikely (Qiu et al., 2019). Still, many surveys relate to an adversarial setting (Wang et al., 2019) or postulate the involvement of industry (Li, 2018) or predict the industry's reaction (Liu et al., 2018).

3.2. Empirical AML studies

Approaching industry practitioners' awareness of Adversarial Machine Learning (AML) revealed significant gaps, as shown in Kumar et al. (2020). The interview-based studies with ML practitioners indicate a lack of competence to protect, detect, or respond to attacks against their ML systems. There is a strong need for research focusing on the development lifecycle of ML, including security aspects. The article identified gaps such as the unpreparedness of ML developers and incident responders and a lack of necessary tools. To address these issues, Kumar et al. (2020) laid a research agenda to improve secure machine learning development within the industry context. Several studies approached industry practitioners directly, extracting the causes for the low AML awareness among ML practitioners (Machine Learning Security in Industry, 2022; Mink et al., 2023; Boenisch et al., 2021).

Our study analyzes the current challenges and available solutions, assesses the alignment between academia and industry, and offers suggestions promoting problem-driven research to enhance collaboration between industry and academia.

These presented studies have already significantly contributed to bridging the gap between academia and industry, paving the way for stronger collaborations. However, this study takes a unique approach by investigating the relationship between academia and industry from an academic perspective. It analyzes study attributes and circumstances in the literature published in journals so far through the metrics of Rigor and Relevance. After shedding some light on this topic, we aim to bring academia and industry by revealing and addressing the reasons for this gap. Encouraging collaboration between academia and industry is vital, as this will open up more opportunities and possibilities for research and hopefully lead to secure and efficient solutions in ML-based applications. ML-based applications are increasingly becoming a focal point in our society and are being used daily.

4. Research methodology

We conducted a systematic literature review (SLR) following the guidelines by Kitchenham et al. (2009):

1. Defining the research questions
2. Planning the search process
3. Setting up the inclusion and exclusion criteria
4. Determining the quality assessment
5. Creating a data collection and analysis method

Fig. 4 provides an overview of the research methodology and summarizes the number of studies after every stage.

4.1. Research questions

We seek answers to the following research questions (RQs):

- **RQ1** What challenges exist in the state-of-the-art and state-of-practice concerning threats associated with machine learning-based software-intensive products and services?
- **RQ2** To what extent have solutions suggested in state-of-the-art been based on challenges in state-of-practice and been validated empirically?

RQ1 aims to identify security-related challenges in academia and industry related to threats against machine learning models. RQ2 investigates to what extent the solutions in academic articles address the challenges pointed out by the industry and how these solutions are evaluated.

4.2. Search strategy

Our literature search steps are outlined in Fig. 4 and elaborated on in subsequent sections.

4.2.1. Search string building

We started our search by utilizing the research questions to determine important keywords for studies related to our topic. Then, we combined these keywords with operators and added quotation marks, wildcards, and braces to increase the quality of the results. This string-building phase resulted in the following search string:

Q1: "TITLE-ABS-KEY (advers AND ("machine learning" OR ml) AND industry)"*

Q2: "(SRCTITLE ("IEEE Security and Privacy") OR SRCTITLE ("usenix security") OR SRCTITLE ("Distributed System Security Symposium") OR SRCTITLE ("IEEE Symposium on Security and Privacy") OR SRCTITLE ("ACM Conference On Computer And Communications Security") AND TITLE-ABS-KEY (advers AND ("machine learning" OR ml) AND industry))"*

Q1 looks for studies that include the fragment "advers", the words "machine learning" or forms of its abbreviation "ml", and "industry" at least once in its title, abstract, or keywords. Q2 investigates the same keywords as Q1 in conferences articles in top-tier security conferences, which also have been independently selected as top-tier security conferences in other studies (Burcham et al., 2017; Vrhovec et al., 2021).

4.2.2. Source selection

To select suitable digital libraries to search relevant articles, we compared the results through an initial search with keyword candidates on Google Scholar, Scopus, and Web of Science. The digital libraries Scopus and Web of Science provide more consistent results (Naqvi et al., 2023). We considered journals and top-tier security conferences in our study to ensure high-quality articles. Scopus includes 99.11% of the journals in the Web of Science master list (Singh et al., 2021) and the top-tier security conferences which were used by Burcham et al. (2017) and Vrhovec et al. (2021). Scopus has also been used in other literature studies as the only digital library, delivering promising results (Lonetti et al., 2023). This makes Scopus a good choice for our study, as it covers most of the Software Engineering publications (Garousi and Mäntylä, 2016) and offers advanced search options. With Scopus, we can search for articles based on their title, abstract, or keyword section while excluding the remaining text.

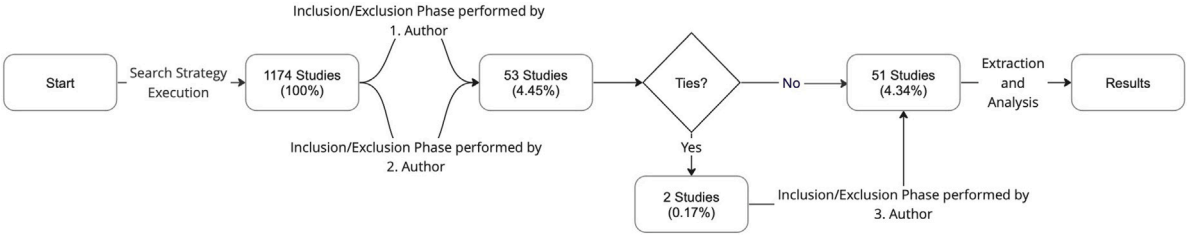

Fig. 4. Research methodology overview.

4.3. Inclusion and exclusion criteria

We have established the following inclusion criteria (IC) list to assess the relevance of articles to our research objective and ensure consistent evaluation among the authors.

- IC1 Manipulation of ML models by a malicious third party
- IC2 Improving ML security
- IC3 Discussing privacy or security attributes of ML models
- IC4 Privacy-Preserving ML models
- IC5 Published in a peer-reviewed journal or top-tier security conference and available as full-text
- IC6 Published before December 2022

IC1 contributes to include studies elaborating on the potential scenario in which a third party attempts to exert its malicious intent by exploiting a vulnerability of an ML model. IC2 and IC3 encapsulate the essential feature of either actively improving the security of ML models or at least discussing it. IC4 singles out our interest in privacy-preserving ML models. IC5 ensures that only peer-reviewed articles that are available as full-text are included. Through IC6, our search results are limited to the search date to ensure the reproducibility of search results.

These inclusion criteria were applied to the title, abstract, and full text for every study in our initial set in the same order. If one of the authors was indecisive about the status of a study during the screening process, the next instance in the order needed to be inspected. For example, if the title and the abstract did not deliver clear information on whether a study would be included or excluded, the author continued with the study's full text.

After discussing the inclusion criteria list, we will explain the corresponding exclusion criteria (EC) list:

- EC1 Adversarial approach to training ML model not for security purposes
- EC2 Applying ML to solve a use case
- EC3 Adversely not used in direct context to an ML model
- EC4 ML not short for Machine Learning
- EC5 Investigating ML models not regarding security
- EC6 Discussing safety issues of ML models
- EC7 Excluding Secondary Study
- EC8 Inversion of criterion IC5

EC1 excludes every article using generative adversarial networks (GANs) since no malicious intent is involved. EC2 excludes all the studies applying ML to contribute to solving real-world scenarios. EC3 removes every article containing the word "adversely" referring to a context different from ML. EC4 removes all the articles in which the abbreviation "ML" refers to concepts other than machine learning. EC5 takes out the studies elaborating on ML models in a context different from security. EC6 removes all the articles discussing the safety issues of ML models, and EC7 excludes all secondary studies. EC8 is an inversion of inclusion criterion IC5.

The articles were filtered by three authors who are experts in the areas of security, machine learning, and software engineering. All studies were screened independently by the first and second authors and compared afterward. In the case of a disagreement, the third author decides if the given study is included or excluded without knowing the decision of the other authors before the assessment.

4.4. Quality assessment

Since journal articles and top-tier security conference articles undergo a rigorous evaluation process by experts in the field, it ensures the reliability and validity of the published studies. However, focusing exclusively on journal articles and top-tier security conference articles as sources itself introduces other types of threats, which we discuss in Section 7 in more detail.

4.5. Analysis methods: Extraction form

We have sorted all the discovered articles using a data extraction form (Kitchenham et al., 2009) and organized them in a spreadsheet. This spreadsheet is made publicly accessible and part of our replicationpackage. Each included article is classified based on four criteria: *Challenges*, *Solutions*, *Rigor*, and *Relevance*. Three authors extract the data from the included articles using the Extraction Form was performed by three authors. In the continuation of this section, we provide details about the four criteria.

4.5.1. Challenges

To address RQ1, we extracted information from the articles regarding the challenges and organized them into three columns. The first column in the spreadsheet, *Challenges (RQ1)*, contains notes about the challenge handled within the respective study. The next column, called *Challenges - Addressed AttackCategories (RQ1)* categorizes the respective study according to the attack classes described by Bieringer et al. (2022), namely, Data Poisoning, Adversarial Examples, and Membership Inference. The categorization will contribute to structuring the challenges in AML research in the context of industry. The column *Challenges - SubCategories (RQ1)* summarizes the notes of the *Challenges (RQ1)* to provide an overview.

4.5.2. Solutions

To answer RQ2, we analyzed the relationship between industry and the suggestions made by the researchers to solve a challenge in a given study. The aim was to record which solutions are provided by academia to support industry in identifying trends and gaps. Furthermore, RQ2 asks how and whether the empirical evaluation of the suggested solutions in the context of industry was conducted.

4.5.3. Rigor

The term "Rigor" generally describes how thorough or whether a research method was applied correctly (Ivarsson and Gorschek, 2011). However, in this study, "Rigor" will encapsulate to what extent the research method and its results are reported in a given study, following the evaluation method developed by Ivarsson and Gorschek (2011). Rigor consists of three aspects: Context, Study Design, and Validity.

Context. This term refers to the surroundings where a study was performed, who was involved, and for whom the results are useful. This Rigor aspect helps to compare a given study to another one.

Study design. Study design describes and explains the setup of a study and supports the reader in comprehending every single decision made regarding the setup.

Validity. The Validity score reflects how thoroughly a study's limitations are reported. It helps the reader to interpret the results of a study.

Our Extraction Form captures the overall Rigor (score) and its three aspects: Context Description, Study Design Description, and Validity (Ivarsson and Gorschek, 2011). Each column title originated from one of the Rigor's aspects and contained the score estimated and discussed by the author team for the respective study.

4.5.4. Relevance

The term "Relevance" encapsulates the potential impact of academic work on industry (Ivarsson and Gorschek, 2011). Ivarsson and Gorschek (2011) propose four aspects to quantify the potential study's impact on the industry: Subject, context, scale, and research method.

Subjects. This aspect of relevance refers to representatives of the sampled audience participating in an empirical evaluation of a proposed solution developed in a study.

Context. To evaluate the usability of a solution proposed in a study, the practitioner must understand the circumstances of a study's proposed solution to determine whether the results are potentially useful in an industrial setting.

Scale. This term reflects to what degree the orders of magnitude used within a study reflect the industrial environment.

Research method. This relevance aspect inspects the potential of an applied research method to deliver potentially useful results for practitioners. For example, applying action research in collaboration with a company to test a new threat modeling guide for software systems incorporating an ML model will likely deliver more useful results for a practitioner than testing it in a laboratory setting on a self-implemented system.

Relevance is a key term in this study because it supports evaluating the included studies' potential impact on industry and identifies gaps that future studies can address by applying the scoring rubric for Relevance proposed by Ivarsson and Gorschek (2011).

With five columns, our Extraction Form features Relevance and its four aspects: Context, Research Method, Subjects, and Scale (Ivarsson and Gorschek, 2011). Each column received the title after one of the aspects of Relevance and incorporated the score we assigned to a given study. We summarized the overall Relevance score in a column called *Overall Score (Relevance)*. The solution analysis was conducted based on the scores of Rigor and Relevance aspects.

5. Results

In this section, we report our results collected during this study. First, we will highlight the course of publications in the last years and the overall Rigor and Relevance score of the included studies. Then, we will dive deeper into the addressed attack categories and their challenges.

5.1. Macro analysis

Fig. 5 shows an overview of publications per year of studies that have been identified in our study. Without defining a time frame during the search process in the inclusion and exclusion criteria, the included studies have been published between 2018 and October 2022. In 2018 and 2019, only one study was published. In contrast, beginning with the year 2020, the published studies increased drastically and peaked at 21 studies in 2022, indicating growing popularity and interest for the intersection of AML and industry.

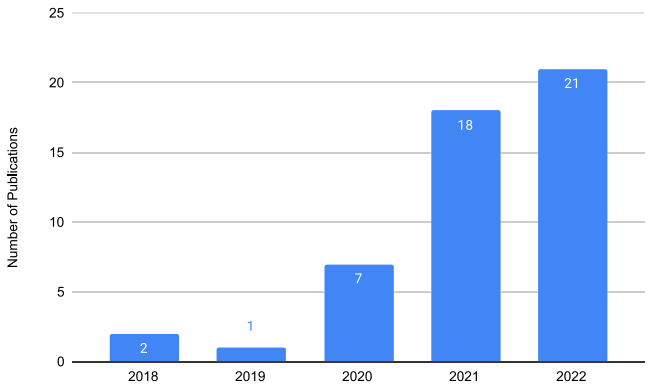

Fig. 5. Number of publications per year.

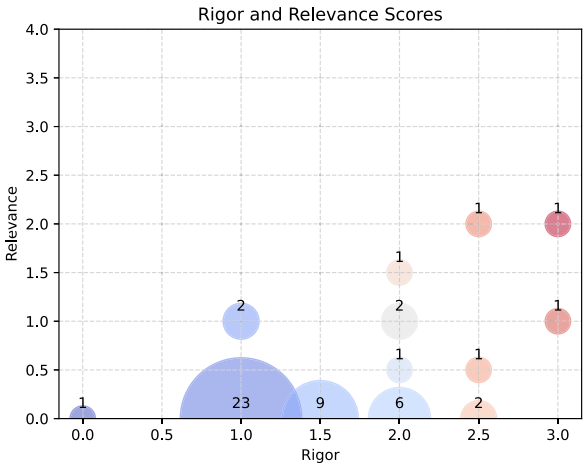

Fig. 6. Rigor and relevance mapping.

Table 2 provides an overview of the included studies' title and solution subcategory each study addressed and the study's achieved overall Rigor and Relevance score.

Fig. 6 displays a bubble chart documenting the distribution of the Relevance and Rigor score each included study received during its assessment , which is also documented in the Extraction Form. The *x*-axis denotes the Rigor score. The *y*-axis covers the Relevance score. The size of every bubble in the chart indicates the number of studies scoring the same overall score for Rigor and Relevance. The center of every bubble represents the number of studies with the same Rigor and Relevance score tuple. For example, the largest bubble in Fig. 6 represents 23 studies that collected in total one point within all the Rigor aspects and zero points within all the Relevance aspects. No study scored the full score since no studies are displayed in the top right corner of Fig. 6. Low Relevance score but a mediocre Rigor score in all extracted studies. Most studies received at least one point overall for their rigor, but a few studies received a higher score. A closer look at the Relevance aspect scores is needed to be reported for further analysis in the next subsection.

5.2. Challenges in the state-of-the-art and state-of-practice concerning threats associated with machine learning-based systems

This part of the study explains and lists the identified challenges and discusses each challenge's rigor and relevance scores. The end of this section contains a challenge analysis.

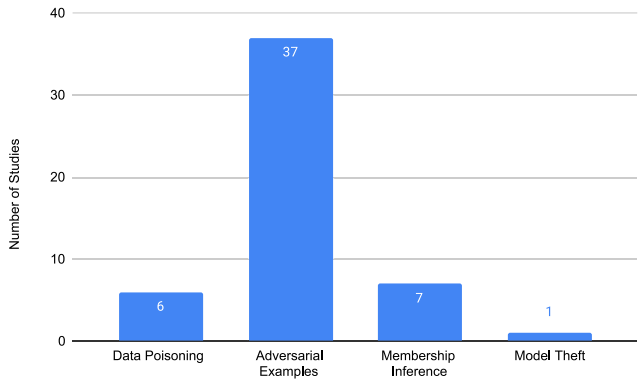

Fig. 7. Addressed attack overview.

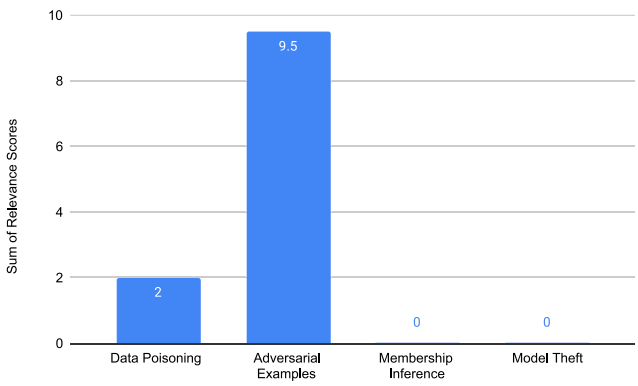

Fig. 8. Overall relevance score for each addressed challenge.

5.2.1. Identified challenges

We summarize the results collected during this study to answer RQ1. Initially, we map the included studies based on the attacks they address since, in the security domain, attacks are a more common term. Then, we inspect the challenges the studies identified and categorize them. We summarize which attack received the most attention and provide an overview of the Relevance scores achieved by every study addressing a certain attack.

Fig. 7 summarizes the distribution of attacks on machine learning models addressed by the included studies, and therefore, it indicates which challenge receives the most attention. 37 studies address Adversarial Examples, six studies Data Poisoning, seven manuscripts address Membership Inference, and one study investigates Model Theft. No study addressed Backdoor attacks.

The low number of studies addressing known attack types, such as Backdoor and Model Theft, seemed controversial. Therefore, we have conducted a separate search to investigate this issue. We have modified the original search string we used for the initial search and applied the new search string to the same pool of studies consisting of journals and top-tier security conferences:

> *Q3: TITLE-ABS-KEY (“Machine Learning” AND (backdoor OR “model theft”) AND industry) AND (LIMIT-TO (DOCTYPE, “ar”))*
>
> *Q4: SRCTITLE (“IEEE Security and Privacy”) OR SRCTITLE (“usenix security”) OR SRCTITLE (“Distributed System Security Symposium”) OR SRCTITLE (“IEEE Symposium on Security and Privacy”) OR SRCTITLE (“ACM Conference On Computer And Communications Security”) AND TITLE-ABS-KEY (“Machine Learning” AND (backdoor OR “model theft”) AND industry)*

The output of both search queries, Q3 and Q4, has confirmed our previous observation; hence, no additional studies exist in the selected top-tier security conferences or journals researchers that address the attack types Backdoor or Model Theft in the context of industry.

Fig. 8 illustrates the score distribution of the Relevance aspects. The Adversarial Example studies scored the highest number of Relevance points, which were 9.5 points among 37 studies. The second highest amount was scored by studies addressing Data Poisoning, with two points among six studies. No Relevance points were scored by the studies addressing Membership Inference and Model Theft within the eight studies we included in this study. To determine the strong and weak aspects of the included studies regarding Rigor and Relevance, we will display the score distribution of the aspects grouped by the addressed attack to observe if a particular challenge pays more attention to a certain aspect.

Next, we created Table 2, categorizing the challenges with the topics Adversarial Examples, Data Poisoning, Membership Inference, and Model Theft, respectively. Since we discovered various challenges, we decided to group them into subcategories to increase the readability of our findings.

Subcategories in the challenge adversarial examples. For the challenge Adversarial Examples, we identified four subcategories: Detection, Attack, Defense, and Evaluation. The largest subcategory focuses on Defense and contains 19 studies out of 36 mainly working on the improvement of ML-based malware detection systems’ robustness (Abusnaina et al., 2021; Kravchik and Shabtai, 2021; Hajaj et al., 2022) to be less prone to adversarial examples or the adjustment of adversarial retraining techniques w.r.t. scalability (Lin et al., 2022) or federated learning (Ibitoye et al., 2022). The second largest group among the challenge main categories is Attack, which includes nine studies. Most studies develop attacks against specific ML systems within an industrial setting (Liu et al., 2021; Usama et al., 2019; Temple et al., 2021; Zhuo and Ge, 2021), or they increase the efficiency of Adversarial Example generation (Demetrio et al., 2021; Chivukula et al., 2020). The third largest group within the topic of Adversarial Examples is Evaluation, containing seven studies. The main two topics handled by those publications are establishing new metrics for robustness (Javanmard and Soltanolkotabi, 2022) and checking the robustness of ML models (Husnoo and Anwar, 2021; Katzir and Elovici, 2018) under different circumstances (Benedick et al., 2021). The remaining studies in the subcategory Detection inspect the Generalization Error (Paul et al., 2022; Fischer, 2020; Sethi and Kantardzic, 2018) and the Low Performance of Adversarial Examples(AE) Generation Rates (Dai and Shu, 2021; Zhuo et al., 2022). For the attack Adversarial Examples, we identified four categories: Defense, Attack, Evaluation, and Detection. The largest subcategory focuses on Defense and contains 19 studies out of 36 mainly working on the improvement of ML-based malware detection systems’ robustness (Abusnaina et al., 2021; Kravchik and Shabtai, 2021; Hajaj et al., 2022) to be less prone to adversarial examples or the adjustment of adversarial retraining techniques w.r.t. scalability (Lin et al., 2022) or federated learning (Ibitoye et al., 2022). The second largest group among the challenge main categories is Attack, which includes nine studies. Most studies develop attacks against specific ML systems within an industrial setting (Liu et al., 2021; Usama et al., 2019; Temple et al., 2021; Zhuo and Ge, 2021), or they increase the efficiency of Adversarial Example generation (Demetrio et al., 2021; Chivukula et al., 2020). The third largest group within the topic of Adversarial Examples is Evaluation, containing seven studies. The main two topics handled by those publications are establishing new metrics for robustness (Javanmard and Soltanolkotabi, 2022) and checking the robustness of ML models (Katzir and Elovici, 2018) under different circumstances (Benedick et al., 2021). The remaining studies in the subcategory Detection inspect the possibility of detecting Adversarial Examples in different settings such as plain neural networks (Kravchik and Shabtai, 2021), IoT malware (Abusnaina et al., 2021), or image processing applications (Datta Gupta et al., 2020).

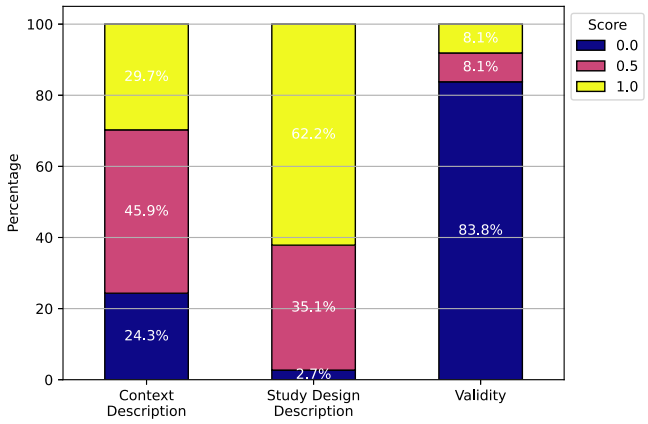

Fig. 9. Overview of rigor aspect scores for adversarial examples studies.

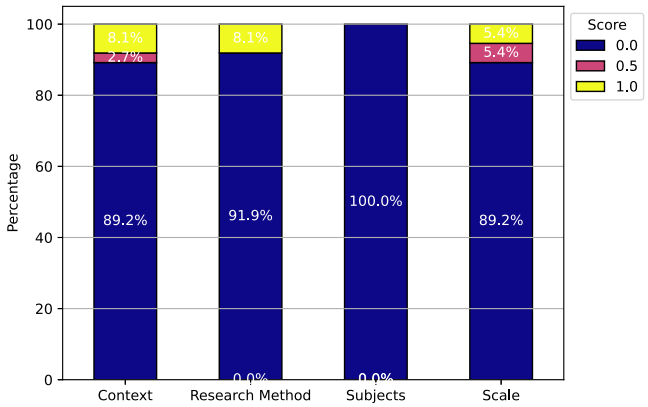

Fig. 10. Overview of relevance aspect scores for adversarial examples studies.

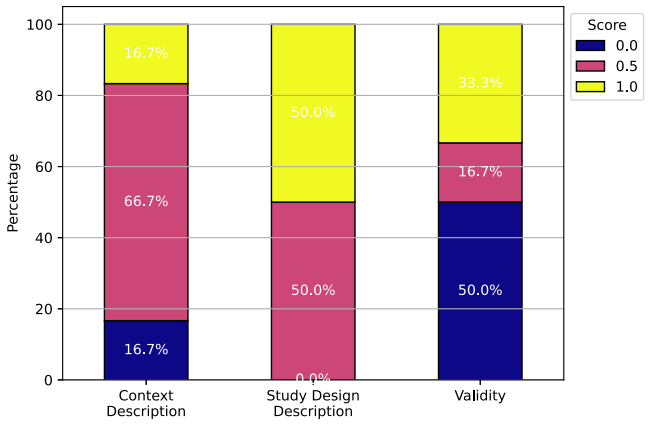

Fig. 11. Overview of rigor aspect scores for data poisoning studies.

Subcategories in the challenges data poisoning, membership inference, and model theft. In the cases of Data Poisoning, Membership Inference, and Model Theft, we observed the main categories of Attack and Defense as shown in Table 2. Out of the seven studies referring to Membership Inference, two studies focus on attack challenges, implementing a dedicated attack tool (Liu et al., 2019) or a new attack variation for Federated Learning (Zhao et al., 2021).

The Defense challenges for Membership Inference vary but center around the resilience improvement of segments in the Federated Learning protocol, suggesting privacy-preserving aggregation of mobile devices (Liu et al., 2022) or privacy-preserving frameworks for industrial Internet of Things (IoT) settings (Arachchige et al., 2020).

In the case of Data Poisoning, we observe a minority of two studies enabling poisoned data detection and correction (Nguyen et al., 2022; Kebande et al., 2021), whereas the remaining four demonstrated attacks on an industrial control system (Kravchik et al., 2022), feature selection process (Liu and Ditzler, 2021), prominent ML algorithms (Yerlikaya and Bahtiyar, 2022) or novel graph-structured prediction models (Xian et al., 2021). These studies focus more on attack-related challenges than defense-related ones.

The only included study addressing Model Theft introduces a defense mechanism called "Proof-of-Learning" which allows the avoidance of false claims of model ownership by adversaries where the model has been trained by multiple parties (Jia et al., 2021).

5.2.2. Rigor and relevance within adversarial examples studies

Fig. 9 shows an overview of the Rigor aspect distribution of studies addressing adversarial examples. The aspect Study Design Description where 62.2 percent of the studies achieved the highest score and 35.1 the second highest score being 1.0 and 0.5, respectively. Also, high scores are denoted by the aspect Context Description. 75.6 percent of the studies scored either 1.0 or 0.5 points. Still, 24.3 percent of the included studies missed the opportunity to provide context for their studies. In the validity aspect, 86.11 percent of the studies scored 0 points, whereas 8.1 percent of the studies received the full score, and 8.1 percent received a score of 0.5.

Fig. 10 illustrates the score distribution of Relevance aspects of studies addressing Adversarial Examples. In the Relevance aspects, Context, and Scale, 89.2% of the studies addressing adversarial examples received a score of 0. 100 and 91.9 percent of the included studies received zero points in the Subject and Research Method sections, respectively. Most of the included studies are laboratory experiments, as shown in Fig. 17, and their results are difficult to transfer to an industrial context. The studies and their experiments are mostly performed in a laboratory setting without a reference for use cases in industry indicated by an overall low Context score. The scale of the data and the data itself used in the included studies do not reflect industrial dimensions. The subjects involved in the experiments had no industrial background, resulting in all studies receiving a score of zero on the Subject rubric.

5.2.3. Rigor and relevance within data poisoning studies

Fig. 11 summarizes the scores of the Rigor aspects for Data Poisoning studies. In the aspect Study Design Description, half of the studies scored the full points, and the other half received 0.5 points. No study received a score of zero in this aspect. For Context Description, 16.67 percent of the studies received the full score, and the majority of studies, 67%, scored 0.5 points. In the validity section, half of the study scored 0 points. One-third of the studies received the full score, and 16.67 percent received a score of 0.5.

In Fig. 12, we documented the Relevance aspect scores for Data Poisoning studies. No study scored in the Research Method aspect because all of them conducted experiments in the lab. In the Context aspect, only 17% of the included studies scored 0.5. The remaining studies received 0 points. For the aspect Scale, scores of 1.0 and 0.5 occurred each in 16.67 percent of the included studies. 67% of the studies did not score in this aspect. In the last aspect, Subject, 100% of the studies received a score of 0.

5.2.4. Rigor and relevance within membership inference studies

In the remaining challenge, Membership Inference, the Rigor and Relevance scores are documented in Fig. 13. Similarly to the other addressed challenges, the Study Design section receives the lowest number of studies that score zero points. 57% of the studies scored the full amount of points, followed by 43% of the studies scoring 0.5 points in this Rigor aspect. The aspect Context Description has an increasing

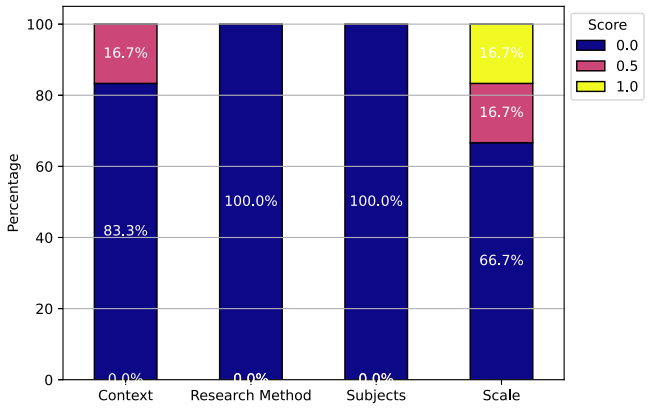

Fig. 12. Overview of relevance aspect scores for data poisoning studies.

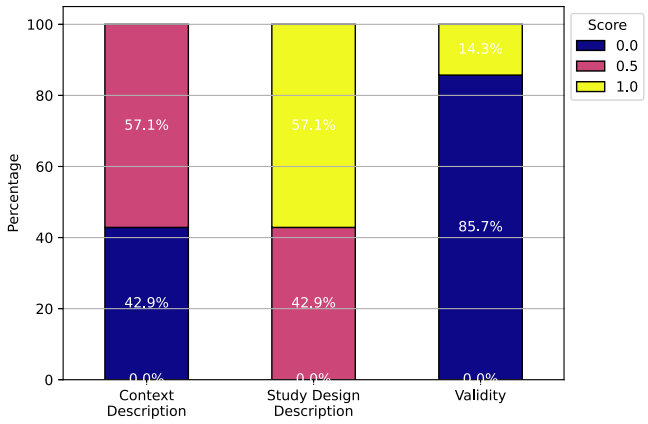

Fig. 13. Overview of rigor aspect scores for membership inference studies.

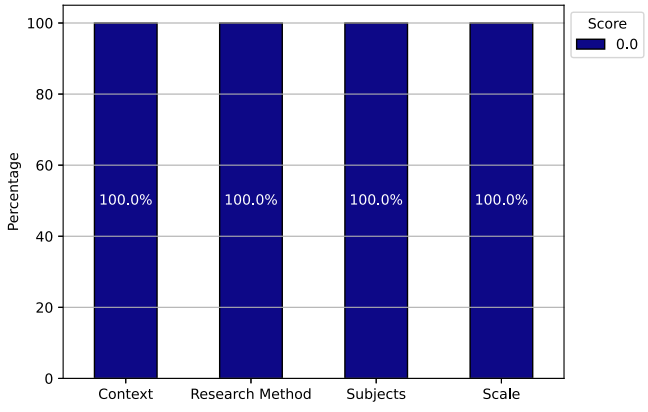

Fig. 14. Overview of relevance aspect scores for membership inference studies.

rate of studies scoring 0 points with 43%. There are no studies with a score of 1.0, but most studies received a 0.5 score. Again, the validity section includes the highest number of studies receiving a score of 0. This is also observable in the distribution of points in the validity section in the other attack categories. Fig. 14 displays that every study received in every aspect of Relevance a score of zero points.

5.2.5. Rigor and relevance within model theft studies

In Fig. 15 summarizes the Rigor aspect scores of the one study addressing Model Theft, which was included in this work. Same as for the studies investigating the challenges Adversarial Examples, Data Poisoning, and Membership Inference, the Context Description and

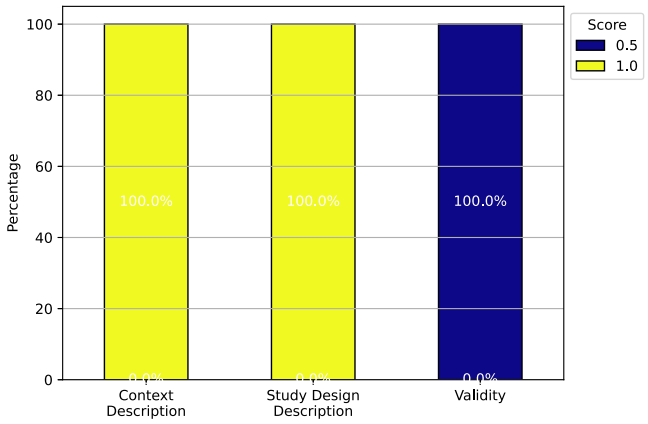

Fig. 15. Overview of rigor aspect scores for model theft studies.

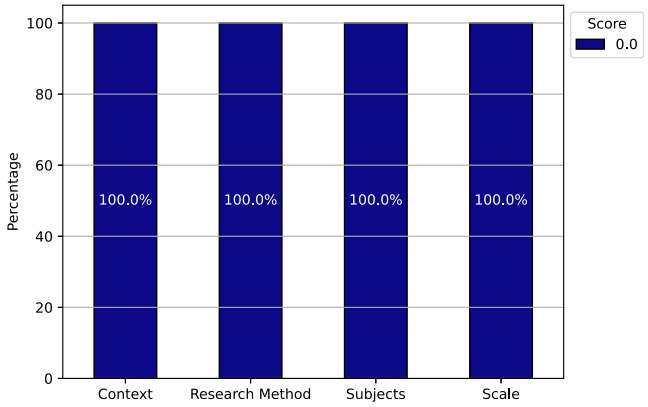

Fig. 16. Overview of relevance aspect scores for model theft studies.

the Study Design Description are extensively reported in the included Model Theft study, whereas, the validity section or any allusion to validity is missing. Fig. 16 depicts the Relevance aspect scores of the one study addressing Model Theft, which was included in this work. The results show the same trend as for the studies investigating Membership Inference. The included Model Theft study has not received any score in the Relevance subcategories, indicating no touchpoints with regard to industry.

5.2.6. Challenge analysis

After categorizing the included studies based on the investigated challenges, we noticed that a majority of studies focus on Adversarial Examples and, in general, strongly focused on improving existing or presenting new attack approaches.

Arms race instead of defense strategy offers. Some studies focused on an attack and proposed an improved approach instead of offering a defense strategy. This observation reinforces the common phenomenon of an "arms race" in the security domain Bertino et al. (2020) and Sadeghi et al. (2020), which is already present in AML subdomains like evasion attacks (Yuan et al., 2019) and defense practices on ML-based Adversarial Malware Detection (Chen et al., 2017; Li et al., 2021; Sadeghi et al., 2020).

Thus, the industry's security experts hesitate to catch up without a guaranteed defense option for every attack improvement study. However, we also realized that the included studies use industry as a "buzzword" in the abstract but failed to design their research in a way that is also comprehensible for industry. We illustrate this claim by analyzing the Rigor and Relevance values of the studies covering defense challenges and comparing them to the attack ones. Based

on our observations, we realize the Rigor is high in both categories, especially in the Study Description; however, the Relevance values are low. The studies are thoroughly designed but fail to provide context, reflect their study's validity, and create an adaptable and reusable study for the industry, except for instrumentalizing industry issues for attention.

Adversarial examples research as misalignment of interest between academia and industry. During our work, we noticed the popularity of Adversarial Examples among the included studies. We argue that studies elaborating on Adversarial Examples are as rewarding as the other attack types in AML but entail complicated bureaucratic processes involving legal issues regarding the required data.

Researching Adversarial Examples is possible with publicly available data sets, hence omitting any form of legal ramifications, which also tremendously impacts the time required to conduct and publish a study. Compared with Data Poisoning, Adversarial Examples studies are simpler because only data entries must be generated to trick the target model successfully. In contrast, Data Poisoning posits multiple tampered data entries and access to the training process. Also, there are no studies published on Backdoor attacks in our included study set, which is a specifically crafted version of Data Poisoning (Li et al., 2022).

Since data and its quality tremendously contribute to the performance of an ML model (Jain et al., 2020) and therefore has an essential role in ML, AML researchers must integrate the same training data as the respective industry branch in their studies to add more value and gain more interest.

A majority of studies in our work predominantly address Adversarial Examples. However, Kumar et al. (2020) and Grosse et al. (2022) published in 2020 and 2022, respectively, state Data Poisoning endangers the industry more after interviewing industry practitioners. Consequently, this circumstance contributes to AML researchers' difficulty finding industry partners for case studies. Additionally, we observe seven out of ten studies in the challenge subcategory focusing on attack realization or optimization.

Strong AML research focus on Rigor. Fig. 6 illustrates the overview of the Rigor and Relevance distribution among the included studies. 98 percent of the included studies scored at least in one Rigor aspect, in contrast, there are 80 percent of the included studies not scoring in any of the Relevance aspects. In Figs. 9, 11, 13 and 15, we observe that the highest scores are reached in the Study Design Description all the addressed challenges by the included studies. Most of the studies are well-documented and explain why a particular step was executed, or a variable was chosen, summing up a high reproducibility of the studies. In the next Rigor aspect, Context Description, we noticed the trend that the number of studies with a score of zero increased compared to the Study Design Description in each addressed challenge. Here, the studies miss elaborating on in which context they are executed and the comparison to other scenarios is challenging. The validity of the studies is rarely mentioned in any included study. Each included study performs poorly in every Relevance aspect, regardless of the addressed challenge. Based on our results, we observe that most studies are thoroughly documented and designed, but as soon as external evaluation criteria like context or validity are taken into account, the scores of the studies drop. The lowest scores are achieved within the Relevance aspects, which analyze the data scale and research method a study applied, the suitability of the subjects participating in the study, and the context or setting in which the study was executed. All categories support the industry in evaluating if the study or parts of it are potentially adaptable or reusable. Since industry pursues topics other than ML threats (Grosse et al., 2022) or is unaware (Kumar et al., 2020), we also identified potential obstacles in the form of legal issues regarding the data sets or disclosing sensitive company information about their infrastructure, concluding a lack of opportunities for collaboration. Additionally, the included studies fail to adopt the setting in a manner that draws the attention of potential industry partners. The proposed additional defense mechanism cannot be incorporated into an industry-hosted system because it is too expensive, the infrastructure cannot uphold it, or it will cause new issues within the system, like bugs or errors. Most studies introduce their work by implying tremendous consequences for industry, frequently applying certain ML algorithms. Still, they fail to incorporate any subjects in their studies who might benefit from the improved security values of a given ML component since most of the studies are experiments empirically evaluating their results instead of choosing research methods involving subjects affected by their research.

Only one study addressed Model Theft (Jia et al., 2021), although it seems to be the most feasible attack. Based on our results, we assume it is challenging to steal a model if it is deployed remotely. On the contrary, studies about Model Theft show promising recovery rates facilitated by attacks on APIs provided by companies like Amazon (Tramèr et al., 2016), or Alibaba and Google offering ML as a Service (MLaaS) (He et al., 2020). A successfully performed Model Theft will lead to tremendous financial consequences and damage the company's reputation irreversibly. Besides training a model is time- and resource-intensive (data collection and cleaning, model fine-tuning, etc.) (Oliynyk et al., 2023), Model Theft can lead to information leaks, money loss, and evasion attacks targeting security applications (Tramèr et al., 2016; Oliynyk et al., 2023).

Most publications we included use publicly available data instead of approaching companies. We assume it is easier for academia to use publicly available data instead of engaging with an industry partner who will impose publication issues regarding data privacy regardless of the type of AML attack. Another approach is to approach government agencies or companies working with publicly available data sets. This commonality is partially motivated by the publication policies of multiple academic venues. This leads to high Rigor values but inhibits industry from relating to the studies and increasing their interest.

Answer to RQ1: Based on our findings from 51 studies, we can summarize the answers to RQ1 as follows:

- Identified challenges with AML of the included studies do not emerge from an industrial context but use industrial applicability of their solutions as motivation
- Adversarial Examples are the most popular researched form of AML attacks, but practitioners are more concerned about Data Poisoning

5.3. Solutions and their empirical validation

Fig. 17 summarizes that 95.9 percent of the included studies are experiments and the remaining ones are case studies. Similarly to the challenge selection process, the researchers did the empirical validation in a lab environment and lacked industry involvement. Therefore, the solutions proposed in the included studies address the challenges discovered by academia in a lab environment.

In the Table 2, the subcategories, *Attack* and *Defense*, are present in each challenge, except in Model Theft which only contains one included study. We listed all the studies only attacking Machine Learning in the *Attack* subcategory. Studies that focus on defense strategies defending against AML attacks are part of the *Defense* subcategory. A large majority of the included studies addressed *Adversarial Examples*, to increase the readability, we introduced besides the subcategories *Attack* and *Defense*, also *Detection* and *Evaluation*. The *Detection* subcategory contains studies searching for specific patterns to determine if Adversarial Examples target an ML model. The subcategory *Evaluation* lists studies assessing the robustness of a given ML system or model against Adversarial Examples. Both subcategories, *Detection* and *Evaluation*, can also be seen as parts of the *Defense* subcategory.

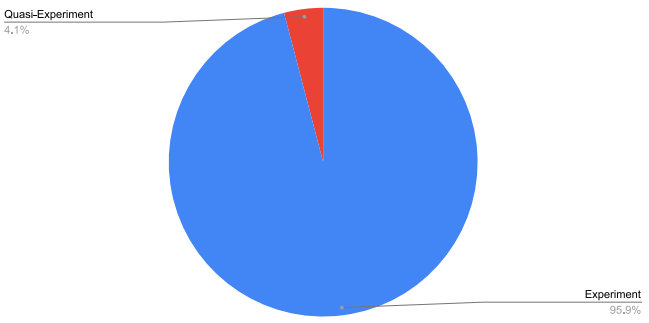

Fig. 17. Study type distribution.

Every subcategory in all challenges scores a higher Rigor score than the Relevance score. In Table 2, we observe Data Poisoning is the only challenge where more studies are listed in the *Attack* than in the *Defense* subcategory. In Table 2, we report a Relevance score of zero for all Membership Inference's subcategories. Although most studies are listed in the Adversarial Examples challenge, we observe low Relevance scores in all the Adversarial Examples subcategories.

5.3.1. Current AML research as starting point for academia and industry collaboration

All challenges and solution propositions are respectively discovered or developed by researchers in a laboratory environment. According to the applied assessment metrics, Rigor, and Relevance, no study fulfilled all the criteria to achieve the full score. Most of the included studies were experiments conducted without industry context, applying an amount and type of data a practitioner could barely relate to. Furthermore, the researchers empirically evaluated the experiments' outcomes without consulting or even considering industry partners. All these points impede companies from comprehending and potentially implementing the content of the studies in their products. Still, most of the researchers introduce their work by alluding to industry products that potentially incorporate the same type of ML to stress the Relevance of their work.

We noticed that most experiments proposed in the included studies are toy examples, which is also a common issue in Software Engineering research for AI-based systems (Martínez-Fernández et al., 2022). However, we realized that new metrics or detection methods suggested in some included studies, as shown in Table 2 are potentially helpful for companies in an ML model's data collection and cleaning phase to detect and correct poisoned data points or during model testing to apply certain robustness metrics, as shown in Table 2. Evaluating those proposed metrics and detection and correction methods in an industrial environment could mutually benefit academia and industry and open new research opportunities. The scalability and applicability of the proposed metrics and methods could play a main role for practitioners, which needs to be included as a key factor in future research. Methods and metrics must be tested on a lower scale to evaluate their effectiveness. Still, no study building upon either a metric or a method tested on a small scale investigated the application in an industrial setting.

5.3.2. Motivation analysis for an industrial context

Table 1 displays the result of a motivation analysis of eight studies from our included studies applying the term "industrial" in their title based on their problem statements and future work. We investigate the problem statements and their references in the introduction section and the sections describing the experimental setup, motivating the industrial setting. This analysis aims to check the origin of problem statements and motivation for an industrial context for which solutions are proposed and to what extent the industry provided input. We want to understand the studies' motivation behind the industrial context, which studies it is built on, and how practitioners and their experiences influence them. Based on the results of our analysis, we attempt to propose new study directions that provide more realistic AML attack and defense scenarios and break out of the current trend of referring to or citing studies discovering issues in an experimental setup.

Table 1
Motivation analysis overview.

Study	Origin of problem statement	Industrial involvement planned in future work
Anthi et al. (2021)	Self claimed	No
Kravchik and Shabtai (2021)	Self claimed	No
Kravchik et al. (2022)	Self claimed	No
Khoda et al. (2019)	Experiment-based	Allusion to finance
Liu et al. (2021)	Experiment-based	No
Benedick et al. (2021)	Experiment-based	No
Zhuo and Ge (2021)	Experiment-based	No
Arachchige et al. (2020)	Experiment-based	No
Ntalampiras (2022)	Related to real-world	No

Self-claimed problem statements. One subset of studies relied on claimed problem statements sourced either by an abstract perspective claiming that the application of ML introduces an additional attack vector to an intrusion detection system (IDS) (Anthi et al., 2021) or Industrial Control System (ICS) (Kravchik and Shabtai, 2021) or building a threat model where an adversary targets an ICS, including an online-trained anomaly detector (Kravchik et al., 2022). We agree on the necessity of those studies because security practitioners need to elicit all the threats that might inflict an AI/ML component in a system. Still, we lack the perspective of involving practitioners who either motivated these studies or can apply the results, to some extent, in their work. Looking at the future work formulated in Kravchik et al. (2022) applying a realistic and practical to simulate an ICS environment using the Tennessee Eastman Process, the two key problems are identified: (1) "(...) improving the attacker's ability to engineer the attack (...)" and (2) "(...) adapting the back-gradient optimization (...)" which, from a practical perspective, misses the consensus that attackers focus on more feasible attack options (Apruzzese et al., 2022) and lack the involvement of practitioners. The same observation applies to the future research directions outlined in Kravchik and Shabtai (2021) and Anthi et al. (2021).

Problem statements based on issues discovered in experiments. In the next study subset, the problem statements are mainly motivated by studies discovering issues of ML models in an experimental setup. More precisely, an approach was designed to select adversarial examples in a manner that guarantees a robust malware defense in Industrial IoT (IIoT) applications using ML (Khoda et al., 2019) as a reaction to the demonstration showing an ineffective way of adversarial retraining based on public data and self-trained model (Grosse et al., 2017). The other study in this subset demonstrates the attack feasibility covered by a white- and black-box approach, but the allusion to real-world applicability is missing because, to the best of our knowledge, there are no instances of Deep Reinforcement Learning (DRL) IIoT systems (Liu et al., 2021). Both studies propose future directions, but only one outlines operational efficiency and financial benefits (Khoda et al., 2019) without mentioning a concrete roadmap of how practitioners can implement the proposed strategies and in which context.

In the last batch of studies, we followed up the citations claiming a wide application of the respective ML model type in industry used to motivate the problem statement and the Relevance for industry.

The systematic robustness evaluation of seven ML and Deep Learning algorithms classifying univariate time-series under perturbation (Benedick et al., 2021) motivated by the fact that previous studies focus more on improving AI systems' accuracy or response time (Dau et al., 2019) without inspecting the training data quality despite numerous studies mention that the data-driven systems' performance

suffers tremendously if the training dataset is of low quality (Lee et al., 2018; Gudivada et al., 2017; Schelter et al., 2018). The focus also remains on data and its protection in the next two studies, suggesting a data protection scheme for industrial monitoring systems (Zhuo and Ge, 2021) and a privacy-preserving framework for industrial IoT systems (Arachchige et al., 2020). Zhuo and Ge (2021) build their problem statement upon the abstract assumption that IoT enlarges the attack surface of industry data and on a non-peer-reviewed (Chakraborty et al., 2018) claiming that current defense mechanisms against Adversarial Examples are mainly model-oriented and a peer-reviewed literature study (Qayyum et al., 2020). However, the results of the second cited systematic literature review inspecting white and gray literature motivated by the fact that MLaaS in third-party cloud servers is one of the many insecure links (Qayyum et al., 2020) and less physically protected than industry sites. Whereas, Arachchige et al. (2020) refers to membership inference and model inversion attacks identified in an experimental setting with publicly available data sets (Abadi et al., 2016) or membership inference attack technique against ML models from "ML as a Service" platforms, Google Prediction API, and Amazon ML, agnostic to the training data and model type (Shokri et al., 2017). Although Shokri et al. (2017) motivates its research by referring to successful membership inference attacks on ML models from "ML as a Service" industrial platforms, it misses the opportunity to incorporate practitioners in its future research intentions and focuses more on improvements on the algorithmic level. This is also the case for the other two studies (Benedick et al., 2021; Zhuo and Ge, 2021) in this batch and propose in their future research directions to dive deeper into the analysis of alternative perturbations or other forms of adversarial training.

The last study of the problem statement analysis investigates adversarial attacks against acoustic monitoring of industrial machines in the medical domain (Ntalampiras, 2022). Same as Shokri et al. (2017), the study cites academic work and insights extracted in an experimental setup facilitating a comparison of Deep Learning models trained on datasets representative of the Internet of Things domain (Li et al., 2017) showing that the application of DL models seems to be promising but is not an argument that DL techniques are widely deployed in IoT. We investigated the citations of Ntalampiras (2022) even further and realized many studies with practical and relevant contributions related to real-world scenarios. One argument for an intensified involvement of AI in safety-critical infrastructure was machine learning is applied to identify faults in a water distribution network serving 3 million people (Ntalampiras, 2014) or in selected use cases in the Finish healthcare sector (Tyrväinen et al., 2018). Also, Ntalampiras (2022) elaborates on the future research directions to extend the efforts towards defense strategies against the discovered attack type and its transferability to other DL models with other architectures, parameters, and training data sets. We classify this study as the one with the closest ties to practical Relevance based on its problem statement and the intentions describing future research directions.

Lack of representative industrial ML systems in the included studies. Furthermore, we investigated in which environment the ML model was embedded in the studies and how the ML models were implemented. We concluded that none(!) of the studies in the industrial problem statement batch used ML model architectures that are also applied in the respective industry domain they were referring to. Also, the environment in which the self-trained ML model was embedded in the experimental setup was self-designed and not evaluated by practitioners of the respective field.

After the problem statement analysis, we suggest improvements to bring practitioners and researchers closer together. Firstly, the validation of the problem statement with practitioners from the respective field, e.g., AML issues in DL-based fault detection system for ICS in a Power Plant, we recommend asking a security expert responsible for power plant applying AI in the same context if the scenario is feasible or representative for the domain. None of the current studies approached practitioners to investigate or validate the problem statement or the proposed solution. Secondly, most of the ML models and the surrounding systems are self-designed and occasionally inspired by a similar model taken from the respective industrial domain.

Included studies alluding to relevance. Here are some example studies conducting research in a real-world industrial setting: In Fredrikson et al. (2014) researchers apply machine learning to identify faults in a water distribution network serving 3 million people (Ntalampiras, 2014). An interpolation poisoning attack is conducted based on real-world ICS data in a realistic test environment for ICS based on the simulated Tennessee Eastman process (Kravchik et al., 2022). Also, instead of publicly available data sets or mocked data, researchers can apply data sets containing data from a representative down-scaled simulation ICS, also known as the Tennesee Eastman Process like in Pan et al. (2015).

We realize that practitioners are overwhelmed by the number of published articles addressing attacks and defense strategies for Machine Learning. This phenomenon is also called an "ML Security Arms Race" by Kästner in his Medium blog.[3] This phenomenon is exemplified in Kravchik et al. (2022) published in 2022 circumventing state-of-the-art adversarial example detection methods (Taormina and Galelli, 2018; Erba et al., 2020; Kravchik and Shabtai, 2021) published in 2018, 2020, and 2021, respectively. In the specific case of Adversarial Examples attacks and the corresponding attribute of model robustness, systematic approaches under more realistic conditions are needed to propose more sustainable defense solutions (Gao et al., 2022) instead of publishing further studies about a new attack variation without proposing a promising defense strategy. Based on the subcategories we propose in our work, the number of studies focusing on attacks on Machine Learning is lower than the number of publications researching defense mechanisms for Machine Learning in every challenge. Additionally, we observe that the proposed solutions among the included studies focus on algorithms and protocols. Still, no publication approaches AML from a Software Architecture perspective and assesses the risk, which would render most of the presented challenges obsolete.

More sources of motivation for future AML studies are available in Database huntr hosted by Protect AI containing bug bounty reports of AI/ML systems[4] or the AI Risk Database hosted by Robust Intelligence[5] or OWASP's Machine Learning Security Top 10.[6]

5.3.3. Issues in deploying solutions proposed by academia

The analysis results for AML-related solutions proposed by academia in the context of industry show three critical issues that need to be addressed by the AML research community to bring Academia and Industry closer together.

Firstly, the motivation to propose a defense solution for a specific AML attack should originate or be present in industry. To this moment, the included studies show that their motivation to propose a solution is based on other studies discovering an AML attack in a laboratory environment.

Secondly, security research should be proactive, thus, the motivation to propose a solution defending an AML laboratory attack is valid. But our motivation analysis in Section 5.3.2 revealed that none of the examined sample studies mentioned in their future work section the testing of their solution in an industrial setting or any other form of industrial collaboration.

[3] https://ckaestne.medium.com/security-and-privacy-in-ml-enabled-systems-1855f561b894.

[4] https://huntr.com/bounties/hacktivity/.

[5] https://airisk.io/.

[6] https://owasp.org/www-project-machine-learning-security-top-10/.

Thirdly, Table 2 shows that the overall relevance score of all(!) Data Poisoning and Membership Inference studies in the Defense subcategory is zero. A lack of relevance indicates that the involved research subjects, the research method, the context and the scale of the studies are misaligned from an industrial perspective, leading to no touch points between industry and academia.

Answer to RQ2:

- Proposed solutions of the included AML studies mainly concern attacking or defending the ML model, contributing to the ML security arms race.
- All empirical evaluations of the identified studies were performed in an experimental setup, limiting the industrial applicability of the solutions.
- The motivation analysis shows that the solutions provided in the studies are based on challenges detected in a laboratory environment without practitioner involvement.
- Proposed future work of the included studies does not indicate that study results will be evaluated in an industrial environment or evaluated by practitioners.

6. Discussion

The results of our study show that current literature conducting research in the intersection between academia and industry is neither motivated by practitioners' experience, designed in a manner that will be reapplied in industry, nor intended to involve practitioners at any point according to our Rigor and Relevance analysis. Therefore, we encourage researchers to approach practitioners to validate their problem statements, experimental setup, and the assumed environment in which the ML model is embedded and evaluate the results. Our results show that if future studies remain on the current course and only refer to attacks discovered or defense strategies tested in an experimental environment, discover new ones in the same setting without approaching at any point practitioners. We compared our results with the ten observations of the "Snapshot of Adversarial Ml Research" section presented in a position study (Apruzzese et al., 2022), which conducted a "systematic paper analysis" on "top" venue conference studies.

Observation #2 made by Apruzzese et al. (2022) states that "Academia and industry perceive adversarial ML differently" whereas our results show that our included studies have their own perception of potential attacks on ML systems used in an industrial environment and propose defense strategies accordingly. "#5 Evidence of adversarial examples in the wild is scarce" is the first observation that fits into the context of our work. It contradicts one of our results, reflecting that adversarial examples receive the most attention among researchers. According to Kumar et al. (2020), practitioners are more concerned about data poisoning than adversarial examples. This shows one misalignment of academia and industry in the context of adversarial ML. The observations #7 and #8 can be summarized as infeasible assumptions regarding the attacker, which we also align with our analysis results. According to our results, a lack of practitioner involvement explains an unrealistic attacker profile and attack circumstances. When practitioners realize the feasibility of an attack on ML systems, they classify it as critical (Machine Learning Security in Industry, 2022). Therefore, we identified the great potential for academia and industry to intertwine when researchers and practitioners collaborate in studies starting from the problem statement or considering it in future research directions after a study is completed, aligning with the recommendation made in Apruzzese et al. (2022).

Despite the popularity of Large Language and Diffusion Models when this study was conducted, we only found and included one study (Khan et al., 2022) investigating the attention-based robustness of Google's Bidirectional Encoder Representations from Transformers (BERT) Large Language Model but we presume that the research activity including Large Language and Diffusion Models will rapidly increase in the near future.

We realize that adversarial ML is a young area with around ten years of research. Still, we see a pattern that researchers tend to dive deeper into the challenges they identified and miss the opportunity to take the turn to practitioners and increase the Relevance of their studies. However, the fact that AI applications are becoming increasingly popular, which can be observed by the growing number of AI applications and ML models on Hugging Face,[7] imposes the urgent need to conduct more relevant adversarial ML research with more realistic settings intensely involving practitioners. Most studies assume the involvement of ML is embedded in an environment of critical infrastructure systems, which are subjects of intensive security audits and strict regulations. Still, research needs to be done on active ML systems in other settings that are not under strict regulations and are not required to have AML defense strategies or conduct an ML threat assessment, rendering a system an attractive target for attackers.

7. Threats to validity

We followed Ampatzoglou et al. (2019) checklist to address threats to validity for secondary studies in software engineering categorized in study selection, data, and research validity.

Study selection validity. The study selection validity refers to all concerns during the search process and the study filtering process (Ampatzoglou et al., 2019). The main threat to the study selection validity is the choice only to consider peer-reviewed studies available in the Scopus digital library. Therefore, it is possible that the authors overlooked relevant articles if they were published in other venues or were only available in other digital libraries. However, the authors specifically searched for peer-reviewed journal and conference articles due to the extensive description of the study design, which is essential to determine the Rigor and Relevance scores. Further, journal articles have been overseen in other AML literature studies (Apruzzese et al., 2022; Arp et al., 2022). Since Scopus includes 99.11% of the journal articles in the Web of Science master list (Singh et al., 2021), the likelihood of missing relevant articles is low.

We formulated our search string as broadly as possible to include all relevant studies using *adversarial* as the main keyword together with *machine learning*. We also tested a more complex search string that included terms like *evasion attack*, *poisoning attack*, *empirical*, and *generative adversarial network*. However, comparing the simpler and more complex search strings by randomly selecting 10% of the results showed that the more complex search string yielded no additional studies. Additionally, during the motivation analysis for industrial context in Section 5.3.2, we performed snowballing on the references in the sections of the included studies describing the problem statement. Since the results of this analysis have shown no involvement of practitioners, a more extensive snowballing process was unpromising.

Another potential reason for overlooking studies is when articles are not written in a language that the authors (English, Swedish, and German) can understand or if the articles are not available in full-text (inclusion criteria IC5).

Threats to the selection validity may arise during the inclusion and exclusion phases due to inadequate execution. A protocol has been established to address this threat, and the involved researchers have documented and discussed decisions. If the inclusion or exclusion status of a given study has been unclear, the researchers have started a voting procedure.

Data validity. Data validity concerns threats to the validity of the extracted dataset and its analysis, which can be identified during the data extraction and analysis phase (Ampatzoglou et al., 2019).

[7] https://huggingface.co/.

Table 2
Overview of AML challenges in industry.

Study	Subcategory	Title	Rigor	Relevance
Adversarial examples			Σ53.5	Σ9.5
Zhuo et al. (2022)	Attack	Attack and defense: Adversarial security of data-driven FDC systems	1	0
Usama et al. (2019)	Attack	The adversarial machine learning conundrum: can the insecurity of ML become the achilles' heel of cognitive networks?	1.5	0
Ntalampiras (2022)	Attack	Adversarial attacks against acoustic monitoring of industrial machines	1	0
Chivukula et al. (2020)	Attack	Game theoretical adversarial deep learning with variational adversaries	1	0
Temple et al. (2021)	Attack	Empirical assessment of generating adversarial configurations for software product lines	1	0
Demetrio et al. (2021)	Attack	Functionality-preserving black-box optimization of adversarial windows malware	1	0
Anthi et al. (2021)	Attack	Adversarial attacks on machine learning cybersecurity defenses in industrial control systems	2.5	0
Qu et al. (2021)	Attack	Frame-correlation transfers trigger economical attacks on deep reinforcement learning policies	2	1
Sethi and Kantardzic (2018)	Attack	Data driven exploratory attacks on black box classifiers in adversarial domains	2	1
Catak et al. (2022)	Defense	Security Hardening of Intelligent Reflecting Surfaces Against Adversarial Machine Learning Attacks	2	0.5
Picot et al. (2022)	Defense	Adversarial robustness via fisher-rao regularization	2.5	0
Dai and Shu (2021)	Defense	Fast-uap: An algorithm for expediting universal adversarial perturbation generation using the orientations of perturbation vectors	1.5	0
Zhuo et al. (2022)	Defense	Attack and defense: Adversarial security of data-driven FDC systems	1	0
Wang et al. (2021)	Defense	Defending adversarial attacks via semantic feature manipulation	1.5	0
Usama et al. (2019)	Defense	The adversarial machine learning conundrum: can the insecurity of ML become the achilles' heel of cognitive networks?	1.5	0
Rossolini et al. (2022)	Defense	Increasing the confidence of deep neural networks by coverage analysis	1	0
Li and Chai (2022)	Defense	Assessing and Enhancing Adversarial Robustness of Predictive Analytics: An Empirically Tested Design Framework	1	0
Liu et al. (2021)	Defense	On deep reinforcement learning security for Industrial Internet of Things	1.5	0
Ibitoye et al. (2022)	Defense	Differentially private self-normalizing neural networks for adversarial robustness in federated learning	1	0
Husnoo and Anwar (2021)	Defense	Do not get fooled: defense against the one-pixel attack to protect IoT-enabled deep learning systems	1.5	0
Nowroozi et al. (2022)	Defense	Demystifying the transferability of adversarial attacks in computer networks	2	0
Hajaj et al. (2022)	Defense	Less is more: Robust and novel features for malicious domain detection	1.5	0
Lin et al. (2022)	Defense	Secure machine learning against adversarial samples at test time	2	0
Paul et al. (2022)	Defense	Ownership recommendation via iterative adversarial training	1	1
Anthi et al. (2021)	Defense	Adversarial attacks on machine learning cybersecurity defenses in industrial control systems	2.5	2
Parulian et al. (2020)	Defense	Effectiveness of the Execution and Prevention of Metric-Based Adversarial Attacks on Social Network Data	1.5	0
Sotgiu et al. (2020)	Defense	Deep neural rejection against adversarial examples	1	0
Khoda et al. (2019)	Defense	Robust malware defense in industrial IoT applications using machine learning with selective adversarial samples	1.5	0
Aghakhani et al. (2020)	Defense	When malware is packin'heat; limits of machine learning classifiers based on static analysis features	2.5	0.5
Abusnaina et al. (2021)	Detection	Dl-fhmc: Deep learning-based fine-grained hierarchical learning approach for robust malware classification	1	0
Kravchik and Shabtai (2021)	Detection	Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca	2	0
Zhuo and Ge (2021)	Detection	Data Guardian: A Data Protection Scheme for Industrial Monitoring Systems	2	1.5
Datta Gupta et al. (2020)	Detection	Determining Sequence of Image Processing Technique (IPT) to Detect Adversarial Attacks	1	0
Maiorca et al. (2020)	Evaluation	Adversarial detection of flash malware: Limitations and open issues	1.5	0
Benedick et al. (2021)	Evaluation	A systematic approach for evaluating artificial intelligence models in industrial settings	3	2
Khan et al. (2022)	Evaluation	BERT Probe: A python package for probing attention based robustness evaluation of BERT models	0	0
Noack et al. (2021)	Evaluation	An empirical study on the relation between network interpretability and adversarial robustness	1	0
Javanmard and Soltanolkotabi (2022)	Evaluation	Precise statistical analysis of classification accuracies for adversarial training	1	0
Fischer (2020)	Evaluation	The conditional entropy bottleneck	1	0
Katzir and Elovici (2018)	Evaluation	Quantifying the resilience of machine learning classifiers used for cybersecurity	1	0
Data poisoning			Σ10	Σ2
Liu and Ditzler (2021)	Attack	Data poisoning against information-theoretic feature selection	1	0
Yerlikaya and Bahtiyar (2022)	Attack	Data poisoning attacks against machine learning algorithms	1	0
Kravchik et al. (2022)	Attack	Practical Evaluation of Poisoning Attacks on Online Anomaly Detectors in Industrial Control Systems	1	1
Xian et al. (2021)	Attack	DeepEC: Adversarial attacks against graph structure prediction models	3	1
Nguyen et al. (2022)	Defense	Effects of image processing operations on adversarial noise and their use in detecting and correcting adversarial images	2	0
Kebande et al. (2021)	Defense	Active machine learning adversarial attack detection in the user feedback process	2	0
Membership inference			Σ8.5	Σ0
Liu et al. (2019)	Attack	Socinf: Membership inference attacks on social media health data with machine learning	1.5	0
Zhao et al. (2021)	Attack	User-Level Membership Inference for Federated Learning in Wireless Network Environment	1	0
Truex et al. (2019)	Defense	Demystifying membership inference attacks in machine learning as a service	1	0
Arachchige et al. (2020)	Defense	A trustworthy privacy preserving framework for machine learning in industrial IoT systems	1	0
Liu et al. (2022)	Defense	Efficient dropout-resilient aggregation for privacy-preserving machine learning	1	0
Nikolaidis et al. (2021)	Defense	Learning realistic patterns from visually unrealistic stimuli: Generalization and data anonymization	2	0
Zhang et al. (2022)	Defense	G-vcfl: grouped verifiable chained privacy-preserving federated learning	1	0
Model theft			Σ2.5	Σ0
Jia et al. (2021)	Defense	Proof-of-learning: Definitions and practice	2.5	0

If a single person is responsible for the data extraction process, it introduces bias due to their subjectivity. To avoid a data extraction bias, three authors performed the data extraction process and applied a data extraction form (available in the replication package). This approach also helps to prevent researcher fatigue, which is a common threat in systematic literature reviews, where the analysis of a large data set can become less rigorous towards the end of the analysis. In cases where the inclusion or exclusion of studies was unclear, decisions were made based on a majority vote by the three authors responsible for data extraction.

A data validity threat arises when the size of the final set of studies is insufficient to draw valid conclusions. This threat is addressed by the very broadly formulated search string.

Another data validity threat refers to the extent to which the quality of studies guarantees the validity of the extracted data. The mitigation for this threat is that we only considered Scopus journals that follow high-quality standards.

During the selection of the classification schema, there is also a threat to data validity. We resolved this issue in our study by using a challenge classification schema based on the ML attack categories proposed by Bieringer et al. (2022).

Research validity. The research validity entails all concerns regarding the overall research design of a study (Ampatzoglou et al., 2019). To ensure that the findings of this study can be replicated, an extraction form was utilized. This form was reviewed and applied by three researchers. The Extraction form has guided the data collection, and the collected data in this study are publicly available as a replication package in an online repository [Link to an online repository containing Extraction Form and Collected Data].

8. Conclusion

ML is not a closed-world problem-solving task anymore and is the subject of growing popularity, attracting sooner or later adversaries utilizing ML to target systems of particular interest. Our study shows that AML studies are immature because they fail to reflect a real-world environment and solve practitioners' security needs. The results of this study report the current degree of maturity of different AML research streams. As a next step, we need to forge comprehendible ML threat assessment methods for practitioners to identify all threats and determine the adequate defense strategy to protect the respective ML system. We aim to establish a new research movement that will actively engage and involve practitioners in adversarial ML studies to prepare current and future ML systems and their stakeholders to defend against threats they will face in realistic scenarios.

CRediT authorship contribution statement

Felix Viktor Jedrzejewski: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Lukas Thode:** Writing – original draft, Data curation, Conceptualization. **Jannik Fischbach:** Writing – review & editing, Data curation. **Tony Gorschek:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition. **Daniel Mendez:** Writing – review & editing. **Niklas Lavesson:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets generated during and/or analyzed during the current study are available at the following URL: https://zenodo.org/records/10654103 (Jedrzejewski, 2024).

Acknowledgments

We would like to acknowledge that this work was supported by the KKS foundation through the SERT Research Profile project (research profile grant 2018/010) at Blekinge Institute of Technology.

Appendix. Overview of AML challenges in industry

See Table 2.

References

Abadi, Martin, Chu, Andy, Goodfellow, Ian, McMahan, H Brendan, Mironov, Ilya, Talwar, Kunal, Zhang, Li, 2016. Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318.

Abusnaina, Ahmed, Abuhamad, Mohammed, Alasmary, Hisham, Anwar, Afsah, Jang, Rhongho, Salem, Saeed, Nyang, Daehun, Mohaisen, David, 2021. Dl-fhmc: Deep learning-based fine-grained hierarchical learning approach for robust malware classification. IEEE Trans. Dependable Secure Comput. 19 (5), 3432–3447.

Aghakhani, Hojjat, Gritti, Fabio, Mecca, Francesco, Lindorfer, Martina, Ortolani, Stefano, Balzarotti, Davide, Vigna, Giovanni, Kruegel, Christopher, 2020. When malware is packin'heat; limits of machine learning classifiers based on static analysis features. In: Network and Distributed Systems Security (NDSS) Symposium 2020.

Ampatzoglou, Apostolos, Bibi, Stamatia, Avgeriou, Paris, Verbeek, Marijn, Chatzigeorgiou, Alexander, 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Inf. Softw. Technol. 106, 201–230.

Anderson, Hyrum, 2021. The practical divide between adversarial ML research and security practice: A red team perspective. USENIX Enigma.

Anthi, Eirini, Williams, Lowri, Rhode, Matilda, Burnap, Pete, Wedgbury, Adam, 2021. Adversarial attacks on machine learning cybersecurity defences in industrial control systems. J. Inf. Secur. Appl. 58, 102717.

Apruzzese, Giovanni, Anderson, Hyrum S., Dambra, Savino, Freeman, David, Pierazzi, Fabio, Roundy, Kevin A., 2022. "Real attackers don't compute gradients": Bridging the gap between adversarial ML research and practice. http://dx.doi.org/10.48550/ARXIV.2212.14315.

Arachchige, Pathum Chamikara Mahawaga, Bertok, Peter, Khalil, Ibrahim, Liu, Dongxi, Camtepe, Seyit, Atiquzzaman, Mohammed, 2020. A trustworthy privacy preserving framework for machine learning in industrial IoT systems. IEEE Trans. Ind. Inform. 16 (9), 6092–6102.

Arp, Daniel, Quiring, Erwin, Pendlebury, Feargus, Warnecke, Alexander, Pierazzi, Fabio, Wressnegger, Christian, Cavallaro, Lorenzo, Rieck, Konrad, 2022. Dos and don'ts of machine learning in computer security. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 3971–3988.

Barreno, Marco, Nelson, Blaine, Sears, Russell, Joseph, Anthony D, Tygar, J Doug, 2006. Can machine learning be secure? In: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. pp. 16–25.

Benedick, Paul-Lou, Robert, Jérémy, Le Traon, Yves, 2021. A systematic approach for evaluating artificial intelligence models in industrial settings. Sensors 21 (18), 6195.

Bertino, Elisa, Singhal, Anoop, Srinivasagopalan, Srivathsan, Verma, Rakesh, 2020. Developing a compelling vision for winning the cybersecurity arms race. In: Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy. pp. 220–222.

Bertolini, Massimo, Mezzogori, Davide, Neroni, Mattia, Zammori, Francesco, 2021. Machine learning for industrial applications: A comprehensive literature review. Expert Syst. Appl. 175, 114820.

Bieringer, Lukas, Grosse, Kathrin, Backes, Michael, Biggio, Battista, Krombholz, Katharina, 2022. Industrial practitioners' mental models of adversarial machine learning. In: Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022). pp. 97–116.

Biggio, Battista, Roli, Fabio, 2018. Wild patterns: Ten years after the rise of adversarial machine learning. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. pp. 2154–2156.

Bilge, Leyla, Dumitraş, Tudor, 2012. Before we knew it: an empirical study of zero-day attacks in the real world. In: Proceedings of the 2012 ACM Conference on Computer and Communications Security. pp. 833–844.

Boenisch, Franziska, Battis, Verena, Buchmann, Nicolas, Poikela, Maija, 2021. "I never thought about securing my machine learning systems": A study of security and privacy awareness of machine learning practitioners. In: Proceedings of Mensch Und Computer 2021. pp. 520–546.

Burcham, Morgan, Al-Zyoud, Mahran, Carver, Jeffrey C, Alsaleh, Mohammed, Du, Hongying, Gilani, Fida, Jiang, Jun, Rahman, Akond, Kafalı, Özgür, Al-Shaer, Ehab, et al., 2017. Characterizing scientific reporting in security literature: An analysis of ACM CCS and IEEE S&P papers. In: Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp. pp. 13–23.

Catak, Ferhat Ozgur, Kuzlu, Murat, Tang, Haolin, Catak, Evren, Zhao, Yanxiao, 2022. Security hardening of intelligent reflecting surfaces against adversarial machine learning attacks. IEEE Access 10, 100267–100275.

Chakraborty, Anirban, Alam, Manaar, Dey, Vishal, Chattopadhyay, Anupam, Mukhopadhyay, Debdeep, 2018. Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.

Chen, Lingwei, Ye, Yanfang, Bourlai, Thirimachos, 2017. Adversarial machine learning in malware detection: Arms race between evasion attack and defense. In: 2017 European Intelligence and Security Informatics Conference. EISIC, pp. 99–106. http://dx.doi.org/10.1109/EISIC.2017.21.

Chivukula, Aneesh Sreevallabh, Yang, Xinghao, Liu, Wei, Zhu, Tianqing, Zhou, Wanlei, 2020. Game theoretical adversarial deep learning with variational adversaries. IEEE Trans. Knowl. Data Eng. 33 (11), 3568–3581.

Cinà, Antonio Emanuele, Grosse, Kathrin, Demontis, Ambra, Vascon, Sebastiano, Zellinger, Werner, Moser, Bernhard A., Oprea, Alina, Biggio, Battista, Pelillo, Marcello, Roli, Fabio, 2023. Wild patterns reloaded: A survey of machine learning security against training data poisoning. ACM Comput. Surv. (ISSN: 0360-0300) http://dx.doi.org/10.1145/3585385, Just Accepted.

Dai, Jiazhu, Shu, Le, 2021. Fast-uap: An algorithm for expediting universal adversarial perturbation generation using the orientations of perturbation vectors. Neurocomputing 422, 109–117.

Datta Gupta, Kishor, Akhtar, Zahid, Dasgupta, Dipankar, 2020. Determining sequence of image processing technique (IPT) to detect adversarial attacks. arXiv e-prints, arXiv–2007.

Dau, Hoang Anh, Bagnall, Anthony, Kamgar, Kaveh, Yeh, Chin-Chia Michael, Zhu, Yan, Gharghabi, Shaghayegh, Ratanamahatana, Chotirat Ann, Keogh, Eamonn, 2019. The UCR time series archive. IEEE/CAA J. Autom. Sin. 6 (6), 1293–1305.

Demetrio, Luca, Biggio, Battista, Lagorio, Giovanni, Roli, Fabio, Armando, Alessandro, 2021. Functionality-preserving black-box optimization of adversarial windows malware. IEEE Trans. Inf. Forensics Secur. 16, 3469–3478.

Erba, Alessandro, Taormina, Riccardo, Galelli, Stefano, Pogliani, Marcello, Carminati, Michele, Zanero, Stefano, Tippenhauer, Nils Ole, 2020. Constrained concealment attacks against reconstruction-based anomaly detectors in industrial control systems. In: Annual Computer Security Applications Conference. pp. 480–495.

Fischer, Ian, 2020. The conditional entropy bottleneck. Entropy 22 (9), 999.

Fredrikson, Matt, Jha, Somesh, Ristenpart, Thomas, 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. pp. 1322–1333.

Fredrikson, Matthew, Lantz, Eric, Jha, Somesh, Lin, Simon, Page, David, Ristenpart, Thomas, 2014. Privacy in pharmacogenetics: An $\{End-to-End\}$ case study of personalized warfarin dosing. In: 23rd USENIX Security Symposium (USENIX Security 14). pp. 17–32.

Gao, Lijun, Yan, Zheng, Liang, Xueqin, Xu, Xi, Wang, Jie, Ding, Wenxiu, Yang, Laurence Tianruo, 2022. Taxonomy and recent advance of game theoretical approaches in adversarial machine learning: A survey. ACM Trans. Sensor Netw..

Garousi, Vahid, Mäntylä, Mika V., 2016. Citations, research topics and active countries in software engineering: A bibliometrics study. Comp. Sci. Rev. 19, 56–77.

Goodfellow, Ian J., Shlens, Jonathon, Szegedy, Christian, 2015. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Grosse, Kathrin, Bieringer, Lukas, Besold, Tarek Richard, Biggio, Battista, Krombholz, Katharina, 2022. "Why do so?"–A practical perspective on machine learning security. arXiv preprint arXiv:2207.05164.

Grosse, Kathrin, Papernot, Nicolas, Manoharan, Praveen, Backes, Michael, McDaniel, Patrick, 2017. Adversarial examples for malware detection. In: Computer Security–ESORICS 2017: 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II 22. Springer, pp. 62–79.

Gudivada, Venkat, Apon, Amy, Ding, Junhua, 2017. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. Int. J. Adv. Softw. 10 (1), 1–20.

Hajaj, Chen, Hason, Nitay, Dvir, Amit, 2022. Less is more: Robust and novel features for malicious domain detection. Electronics 11 (6), 969.

He, Yingzhe, Meng, Guozhu, Chen, Kai, Hu, Xingbo, He, Jinwen, 2020. Towards security threats of deep learning systems: A survey. arXiv:1911.12562.

Husnoo, Muhammad Akbar, Anwar, Adnan, 2021. Do not get fooled: defense against the one-pixel attack to protect IoT-enabled deep learning systems. Ad Hoc Netw. 122, 102627.

Ibitoye, Olakunle, Shafiq, M. Omair, Matrawy, Ashraf, 2022. Differentially private self-normalizing neural networks for adversarial robustness in federated learning. Comput. Secur. 116, 102631.

Ivarsson, Martin, Gorschek, Tony, 2011. A method for evaluating rigor and industrial relevance of technology evaluations. Empir. Softw. Eng. 16, 365–395.

Jain, Abhinav, Patel, Hima, Nagalapatti, Lokesh, Gupta, Nitin, Mehta, Sameep, Guttula, Shanmukha, Mujumdar, Shashank, Afzal, Shazia, Sharma Mittal, Ruhi, Munigala, Vitobha, 2020. Overview and importance of data quality for machine learning tasks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3561–3562.

Javanmard, Adel, Soltanolkotabi, Mahdi, 2022. Precise statistical analysis of classification accuracies for adversarial training. Ann. Statist. 50 (4), 2127–2156.

Jedrzejewski, Felix Viktor, 2024. Dataset for Adversarial Machine Learning in Industry: A Systematic Literature Review. http://dx.doi.org/10.5281/zenodo.10654103.

Jia, Hengrui, Yaghini, Mohammad, Choquette-Choo, Christopher A, Dullerud, Natalie, Thudi, Anvith, Chandrasekaran, Varun, Papernot, Nicolas, 2021. Proof-of-learning: Definitions and practice. In: 2021 IEEE Symposium on Security and Privacy. SP, IEEE, pp. 1039–1056.

Katzir, Ziv, Elovici, Yuval, 2018. Quantifying the resilience of machine learning classifiers used for cyber security. Expert Syst. Appl. 92, 419–429.

Kebande, Victor R, Alawadi, Sadi, Awaysheh, Feras M, Persson, Jan A, 2021. Active machine learning adversarial attack detection in the user feedback process. IEEE Access 9, 36908–36923.

Khan, Shahrukh, Shahid, Mahnoor, Singh, Navdeeppal, 2022. BERT Probe: A python package for probing attention based robustness evaluation of BERT models. Softw. Impacts 13, 100310.

Khoda, Mahbub E, Imam, Tasadduq, Kamruzzaman, Joarder, Gondal, Iqbal, Rahman, Ashfaqur, 2019. Robust malware defense in industrial IoT applications using machine learning with selective adversarial samples. IEEE Trans. Ind. Appl. 56 (4), 4415–4424.

Kitchenham, Barbara, Brereton, O Pearl, Budgen, David, Turner, Mark, Bailey, John, Linkman, Stephen, 2009. Systematic literature reviews in software engineering–a systematic literature review. Inf. Softw. Technol. 51 (1), 7–15.

Kravchik, Moshe, Demetrio, Luca, Biggio, Battista, Shabtai, Asaf, 2022. Practical evaluation of poisoning attacks on online anomaly detectors in industrial control systems. Comput. Secur. 122, 102901.

Kravchik, Moshe, Shabtai, Asaf, 2021. Efficient cyber attack detection in industrial control systems using lightweight neural networks and pca. IEEE Trans. Dependable Secure Comput. 19 (4), 2179–2197.

Kumar, Ram Shankar Siva, Nyström, Magnus, Lambert, John, Marshall, Andrew, Goertzel, Mario, Comissoneru, Andi, Swann, Matt, Xia, Sharon, 2020. Adversarial machine learning-industry perspectives. In: 2020 IEEE Security and Privacy Workshops. SPW, IEEE, pp. 69–75.

Lee, Jay, Davari, Hossein, Singh, Jaskaran, Pandhare, Vibhor, 2018. Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. Manuf. Lett. 18, 20–23.

Li, Jian-hua, 2018. Cyber security meets artificial intelligence: a survey. Front. Inf. Technol. Electron. Eng. 19 (12), 1462–1474.

Li, Weifeng, Chai, Yidong, 2022. Assessing and enhancing adversarial robustness of predictive analytics: An empirically tested design framework. J. Manage. Inf. Syst. 39 (2), 542–572.

Li, Peng, Chen, Zhikui, Yang, Laurence Tianruo, Zhang, Qingchen, Deen, M Jamal, 2017. Deep convolutional computation model for feature learning on big data in internet of things. IEEE Trans. Ind. Inform. 14 (2), 790–798.

Li, Yiming, Jiang, Yong, Li, Zhifeng, Xia, Shu-Tao, 2022. Backdoor learning: A survey. IEEE Trans. Neural Netw. Learn. Syst..

Li, Deqiang, Li, Qianmu, Ye, Yanfang, Xu, Shouhuai, 2021. Arms race in adversarial malware detection: A survey. ACM Comput. Surv. 55 (1), 1–35.

Lin, Jing, Njilla, Laurent L., Xiong, Kaiqi, 2022. Secure machine learning against adversarial samples at test time. EURASIP J. Inf. Secur. 2022 (1), 1.

Liu, Heng, Ditzler, Gregory, 2021. Data poisoning against information-theoretic feature selection. Inform. Sci. 573, 396–411.

Liu, Ziyao, Guo, Jiale, Lam, Kwok-Yan, Zhao, Jun, 2022. Efficient dropout-resilient aggregation for privacy-preserving machine learning. IEEE Trans. Inf. Forensics Secur..

Liu, Qiang, Li, Pan, Zhao, Wentao, Cai, Wei, Yu, Shui, Leung, Victor CM, 2018. A survey on security threats and defensive techniques of machine learning: A data driven view. IEEE Access 6, 12103–12117.

Liu, Gaoyang, Wang, Chen, Peng, Kai, Huang, Haojun, Li, Yutong, Cheng, Wenqing, 2019. Socinf: Membership inference attacks on social media health data with machine learning. IEEE Trans. Comput. Soc. Syst. 6 (5), 907–921.

Liu, Xing, Yu, Wei, Liang, Fan, Griffith, David, Golmie, Nada, 2021. On deep reinforcement learning security for industrial internet of things. Comput. Commun. 168, 20–32.

Lonetti, Francesca, Bertolino, Antonia, Di Giandomenico, Felicita, 2023. Model-based security testing in IoT systems: A rapid review. Inf. Softw. Technol. 107326.

2022. Machine learning security in industry: A quantitative survey. IEEE Trans. Inf. Forensics Secur. 18, 1749–1762. http://dx.doi.org/10.1109/TIFS.2023.3251842.

Maiorca, Davide, Demontis, Ambra, Biggio, Battista, Roli, Fabio, Giacinto, Giorgio, 2020. Adversarial detection of flash malware: Limitations and open issues. Comput. Secur. 96, 101901.

Martínez-Fernández, Silverio, Bogner, Justus, Franch, Xavier, Oriol, Marc, Siebert, Julien, Trendowicz, Adam, Vollmer, Anna Maria, Wagner, Stefan, 2022. Software engineering for AI-based systems: a survey. ACM Trans. Softw. Eng. Methodol. (TOSEM) 31 (2), 1–59.

Mink, Jaron, Kaur, Harjot, Schmüser, Juliane, Fahl, Sascha, Acar, Yasemin, 2023. "Security is not my field, I'ma stats guy": A Qualitative Root Cause Analysis of Barriers to Adversarial Machine Learning Defenses in Industry. In: Proc. of USENIX Security.

Naqvi, Bilal, Perova, Kseniia, Farooq, Ali, Makhdoom, Imran, Oyedeji, Shola, Porras, Jari, 2023. Mitigation strategies against the phishing attacks: A systematic literature review. Comput. Secur. 103387.

Nguyen, Huy H, Kuribayashi, Minoru, Yamagishi, Junichi, Echizen, Isao, 2022. Effects of image processing operations on adversarial noise and their use in detecting and correcting adversarial images. IEICE Trans. Inf. Syst. 105 (1), 65–77.

Nikolaidis, Konstantinos, Kristiansen, Stein, Plagemann, Thomas, Goebel, Vera, Liestøl, Knut, Kankanhalli, Mohan, Traaen, Gunn Marit, Overland, Britt, Akre, Harriet, Aakerøy, Lars, et al., 2021. Learning realistic patterns from visually unrealistic stimuli: Generalization and data anonymization. J. Artificial Intelligence Res. 72, 1163–1214.

Noack, Adam, Ahern, Isaac, Dou, Dejing, Li, Boyang, 2021. An empirical study on the relation between network interpretability and adversarial robustness. SN Comput. Sci. 2, 1–13.

Nowroozi, Ehsan, Mekdad, Yassine, Berenjestanaki, Mohammad Hajian, Conti, Mauro, El Fergougui, Abdeslam, 2022. Demystifying the transferability of adversarial attacks in computer networks. IEEE Trans. Netw. Serv. Manag. 19 (3), 3387–3400.

Ntalampiras, Stavros, 2014. Fault identification in distributed sensor networks based on universal probabilistic modeling. IEEE Trans. Neural Netw. Learn. Syst. 26 (9), 1939–1949.

Ntalampiras, Stavros, 2022. Adversarial attacks against acoustic monitoring of industrial machines. IEEE Internet Things J. 10 (4), 2832–2839.

Oliynyk, Daryna, Mayer, Rudolf, Rauber, Andreas, 2023. I know what you trained last summer: A survey on stealing machine learning models and defences. ACM Comput. Surv..

Pan, Shengyi, Morris, Thomas, Adhikari, Uttam, 2015. Classification of disturbances and cyber-attacks in power systems using heterogeneous time-synchronized data. IEEE Trans. Ind. Inform. 11 (3), 650–662.

Papernot, Nicolas, McDaniel, Patrick, Goodfellow, Ian, 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277.

Parulian, Nikolaus Nova, Lu, Tiffany, Mishra, Shubhanshu, Avram, Mihai, Diesner, Jana, 2020. Effectiveness of the execution and prevention of metric-based adversarial attacks on social network data. Information 11 (6), 306.

Paul, Agyemang, Zhao, Xunming, Fang, Luping, Wu, Zhefu, 2022. Ownership recommendation via iterative adversarial training. Neural Process. Lett. 1–19.

Picot, Marine, Messina, Francisco, Boudiaf, Malik, Labeau, Fabrice, Ayed, Ismail Ben, Piantanida, Pablo, 2022. Adversarial robustness via fisher-rao regularization. IEEE Trans. Pattern Anal. Mach. Intell..

Qayyum, Adnan, Ijaz, Aneeqa, Usama, Muhammad, Iqbal, Waleed, Qadir, Junaid, Elkhatib, Yehia, Al-Fuqaha, Ala, 2020. Securing machine learning in the cloud: A systematic review of cloud machine learning security. Front. Big Data 3, 587139.

Qiu, Shilin, Liu, Qihe, Zhou, Shijie, Wu, Chunjiang, 2019. Review of artificial intelligence adversarial attack and defense technologies. Appl. Sci. 9 (5), 909.

Qu, Xinghua, Ong, Yew-Soon, Gupta, Abhishek, 2021. Frame-correlation transfers trigger economical attacks on deep reinforcement learning policies. IEEE Trans. Cybern. 52 (8), 7577–7590.

Reith, Robert Nikolai, Schneider, Thomas, Tkachenko, Oleksandr, 2019. Efficiently stealing your machine learning models. In: Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society. pp. 198–210.

Rossolini, Giulio, Biondi, Alessandro, Buttazzo, Giorgio, 2022. Increasing the confidence of deep neural networks by coverage analysis. IEEE Trans. Softw. Eng. 49 (2), 802–815.

Russell, Stuart J., Norvig, Peter, 2010. Artificial Intelligence a Modern Approach. London.

Sadeghi, Koosha, Banerjee, Ayan, Gupta, Sandeep K.S., 2020. A system-driven taxonomy of attacks and defenses in adversarial machine learning. IEEE Trans. Emerg. Top. Comput. Intell. 4 (4), 450–467.

Schelter, Sebastian, Lange, Dustin, Schmidt, Philipp, Celikel, Meltem, Biessmann, Felix, Grafberger, Andreas, 2018. Automating large-scale data quality verification. Proc. VLDB Endow. 11 (12), 1781–1794.

Sethi, Tegjyot Singh, Kantardzic, Mehmed, 2018. Data driven exploratory attacks on black box classifiers in adversarial domains. Neurocomputing 289, 129–143.

Shokri, Reza, Stronati, Marco, Song, Congzheng, Shmatikov, Vitaly, 2017. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy. SP, IEEE, pp. 3–18.

Singh, Vivek Kumar, Singh, Prashasti, Karmakar, Mousumi, Leta, Jacqueline, Mayr, Philipp, 2021. The journal coverage of web of science, scopus and dimensions: A comparative analysis. Scientometrics 126, 5113–5142.

Sotgiu, Angelo, Demontis, Ambra, Melis, Marco, Biggio, Battista, Fumera, Giorgio, Feng, Xiaoyi, Roli, Fabio, 2020. Deep neural rejection against adversarial examples. EURASIP J. Inf. Secur. 2020, 1–10.

Steinhardt, Jacob, Koh, Pang Wei W., Liang, Percy S., 2017. Certified defenses for data poisoning attacks. Adv. Neural Inf. Process. Syst. 30.

Suciu, Octavian, Marginean, Radu, Kaya, Yigitcan, Daume III, Hal, Dumitras, Tudor, 2018. When does machine learning $\{FAIL\}$? generalized transferability for evasion and poisoning attacks. In: 27th USENIX Security Symposium (USENIX Security 18). pp. 1299–1316.

Taormina, Riccardo, Galelli, Stefano, 2018. Deep-learning approach to the detection and localization of cyber-physical attacks on water distribution systems. J. Water Resour. Plan. Manage. 144 (10), 04018065.

Temple, Paul, Perrouin, Gilles, Acher, Mathieu, Biggio, Battista, Jézéquel, Jean-Marc, Roli, Fabio, 2021. Empirical assessment of generating adversarial configurations for software product lines. Empir. Softw. Eng. 26, 1–49.

Terranova, Nadia, Venkatakrishnan, Karthik, Benincosa, Lisa J., 2021. Application of machine learning in translational medicine: current status and future opportunities. AAPS J. 23 (4), 74.

Tidjon, Lionel Nganyewou, Khomh, Foutse, 2022. Threat assessment in machine learning based systems. arXiv preprint arXiv:2207.00091.

Tramèr, Florian, Zhang, Fan, Juels, Ari, Reiter, Michael K, Ristenpart, Thomas, 2016. Stealing machine learning models via prediction $\{APIs\}$. In: 25th USENIX Security Symposium (USENIX Security 16). pp. 601–618.

Truex, Stacey, Liu, Ling, Gursoy, Mehmet Emre, Yu, Lei, Wei, Wenqi, 2019. Demystifying membership inference attacks in machine learning as a service. IEEE Trans. Serv. Comput. 14 (6), 2073–2089.

Tyrväinen, Pasi, Silvennoinen, Minna, Talvitie-Lamberg, Karoliina, Ala-Kitula, Anniina, Kuoremäki, Reija, 2018. Identifying opportunities for AI applications in healthcare—Renewing the national healthcare and social services. In: 2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH). IEEE, pp. 1–7.

Usama, Muhammad, Qadir, Junaid, Al-Fuqaha, Ala, Hamdi, Mounir, 2019. The adversarial machine learning conundrum: can the insecurity of ML become the achilles' heel of cognitive networks? IEEE Netw. 34 (1), 196–203.

Vrhovec, Simon, Caviglione, Luca, Wendzel, Steffen, 2021. Crème de la crème: Lessons from papers in security publications. In: Proceedings of the 16th International Conference on Availability, Reliability and Security. pp. 1–9.

Wang, Xianmin, Li, Jing, Kuang, Xiaohui, Tan, Yu-an, Li, Jin, 2019. The security of machine learning in an adversarial setting: A survey. J. Parallel Distrib. Comput. 130, 12–23.

Wang, Shuo, Nepal, Surya, Rudolph, Carsten, Grobler, Marthie, Chen, Shangyu, Chen, Tianle, An, Zike, 2021. Defending adversarial attacks via semantic feature manipulation. IEEE Trans. Serv. Comput. 15 (6), 3184–3197.

Wu, Yulei, 2020. Robust learning-enabled intelligence for the internet of things: A survey from the perspectives of noisy data and adversarial examples. IEEE Internet Things J. 8 (12), 9568–9579.

Xian, Xingping, Wu, Tao, Qiao, Shaojie, Wang, Wei, Wang, Chao, Liu, Yanbing, Xu, Guangxia, 2021. DeepEC: Adversarial attacks against graph structure prediction models. Neurocomputing 437, 168–185.

Yerlikaya, Fahri Anıl, Bahtiyar, Şerif, 2022. Data poisoning attacks against machine learning algorithms. Expert Syst. Appl. 208, 118101.

Yuan, Xiaoyong, He, Pan, Zhu, Qile, Li, Xiaolin, 2019. Adversarial examples: Attacks and defenses for deep learning. IEEE Trans. Neural Netw. Learn. Syst. 30 (9), 2805–2824.

Zhang, Jiliang, Li, Chen, 2020. Adversarial examples: Opportunities and challenges. IEEE Trans. Neural Netw. Learn. Syst. 31 (7), 2578–2593. http://dx.doi.org/10.1109/TNNLS.2019.2933524.

Zhang, Zhuangzhuang, Wu, Libing, He, Debiao, Wang, Qian, Wu, Dan, Shi, Xiaochuan, Ma, Chao, 2022. G-vcfl: grouped verifiable chained privacy-preserving federated learning. IEEE Trans. Netw. Serv. Manag. 19 (4), 4219–4231.

Zhao, Yanchao, Chen, Jiale, Zhang, Jiale, Yang, Zilu, Tu, Huawei, Han, Hao, Zhu, Kun, Chen, Bing, 2021. User-level membership inference for federated learning in wireless network environment. Wirel. Commun. Mob. Comput. 2021, 1–17.

Zhuo, Yue, Ge, Zhiqiang, 2021. Data guardian: A data protection scheme for industrial monitoring systems. IEEE Trans. Ind. Inform. 18 (4), 2550–2559.

Zhuo, Yue, Yin, Zhenqin, Ge, Zhiqiang, 2022. Attack and defense: Adversarial security of data-driven FDC systems. IEEE Trans. Ind. Inform. 19 (1), 5–19.

Felix Viktor Jedrzejewski is a second-year Ph.D. student in Software Engineering at the Blekinge Institute of Technology, Sweden. After studying Information Systems at the Technical University Munich. His research focuses on the intersection of Empirical Software Engineering, Machine Learning, and Security with a particular focus on improving the secure development lifecycle of Machine Learning Systems in close collaboration with the relevant industries.

Lukas Thode is a second-year Ph.D. student in Software Engineering at the Blekinge Institute of Technology, Sweden after studying his Bachelor of Arts in Information Systems at the WWU in Munster and his Master of Science in AI Engineering at the Jönköping Tekniska Högskola. His research focuses on the intersection of Empirical Software Engineering and Machine Learning for financial services in close collaboration with the relevant industries.

Dr. Tony Gorschek is a Professor at the Software Engineering Research Lab at Blekinge Institute of Technology (Sweden) and visiting senior researcher at fortiss Germany. He has over fifteen years of industrial experience as a CTO, senior executive consultant and engineer, but also as chief architect and product manager. In addition, he has built up six startups in fields ranging from logistics to Internet based services and algorithmic stock trading. His research interests include empirical software engineering, engineering security, technology and product management, and value based lean development of software intensive products and services. A clear theme in Dr. Gorschek's research is the focus on applied research, where the challenges are based on real industrial needs, and where solutions are developed to address these needs. Most importantly is that research success is measured not only in scientific publications, but more relevant in usability and usefulness of research result in industry through the measurement of efficiency and effectiveness and return on investment.

Jannik Fischbach works as a consultant at Netlight and as a Post-doctoral researcher at fortiss. His research focuses on modern Natural Language Processing techniques and their potential to support stakeholders in the software engineering process. He has already published at several top venues (e.g. ICST, ESEM) and won the Best Industry Paper Award at ESEM'20 and Best Research Paper Award at REFSQ'21.

Daniel Mendez is full professor at the Blekinge Institute of Technology, Sweden, and Lead Researcher heading the research division Requirements Engineering at fortiss, the research and transfer institute of the Free State of Bavaria for software-intensive systems and services. After studying Computer Science and Cognitive Neuroscience at the Ludwig Maximilian University of Munich, he pursued his doctoral and his habilitation degrees at the Technical University of Munich. His research is since then on Empirical Software Engineering with a particular focus on interdisciplinary, qualitative research in Requirements Engineering and its quality improvement — all in close collaboration with the relevant industries. He is further editorial board member for EMSE and JSS where he co-chairs the special tracks Reproducibility & Open Science (EMSE) and In Practice (JSS) respectively. Finally, he is a member of the ACM, the German association of university professors and lecturers, the German Informatics Society, and ISERN.

Niklas Lavesson is Professor of Software Engineering with specialization Applied AI and Machine Learning at Software Engineering Research Laboratory (SERL), Blekinge Institute of Technology (BTH) since July 2021. Lavesson has a Ph.D. in computer science with specialization in machine learning from BTH (2008). His research interests within basic research are focused on explainable and interactive AI-systems. Within applied research, Lavesson is interested in how to (1) use AI and machine learning to make software engineering more efficient and effective, and (2) how to pose requirements for, develop, test, and maintain data-driven AI systems which are able to make themselves understood by humans and which can collaborate with humans to solve tasks together.